Helsinki University of Technology
Dissertations in Computer and Information Science
Espoo 2007 Report D23

METHODS FOR EXPLORING GENOMIC DATA SETS: APPLICATION TO HUMAN ENDOGENOUS RETROVIRUSES

Merja Oja

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Computer Science and Engineering for public examination and debate in Auditorium T2 at Helsinki University of Technology (Espoo, Finland) on the 14th of December, 2007, at 12 o'clock noon.

Helsinki University of Technology
Department of Computer Science and Engineering
Laboratory of Computer and Information Science

Distribution:

Helsinki University of Technology
Laboratory of Computer and Information Science
P.O.Box 5400
FIN-02015 TKK
Finland

Tel. +358-9-451 3272
Fax +358-9-451 3277
http://www.cis.hut.fi/

Available in pdf format at http://lib.hut.fi/Diss/2007/isbn9789512290628/

ISBN 978-951-22-9061-1 (printed version)
ISBN 978-951-22-9062-8 (electronic version)
ISSN 1459-7020

Yliopistopaino Oy
Helsinki 2007

ABSTRACT

Oja, M. (2007): **Methods for exploring genomic data sets: application to human endogenous retroviruses.** Doctoral thesis, Helsinki University of Technology, Dissertations in Computer and Information Science, Report D23, Espoo, Finland.

Keywords: bioinformatics, exploratory data analysis, gene expression, hidden Markov model, human endogenous retrovirus, information visualization, learning metrics, reliability, self-organizing map.

In this thesis exploratory data analysis methods have been developed for analyzing genomic data, in particular human endogenous retrovirus (HERV) sequences and gene expression data. HERVs are remains of ancient retrovirus infections and now reside within the human genome. Little is known about their functions. However, HERVs have been implicated in some diseases. This thesis provides methods for analyzing the properties and expression patterns of HERVs.

Nowadays the genomic data sets are so large that sophisticated data analysis methods are needed in order to uncover interesting structures in the data. The purpose of exploratory methods is to help in generating hypotheses about the properties of the data. For example, by grouping together genes behaving similarly, and hence presumably having similar function, a new function can be suggested for previously uncharacterized genes. The hypotheses generated by exploratory data analysis can be verified later in more detailed studies. In contrast, a detailed analysis of all the genes of an organism would be too time consuming and expensive.

In this thesis self-organizing map (SOM) based exploratory data analysis approaches for visualization and grouping of gene expression profiles and HERV sequences are presented. The SOM-based analysis is complemented with estimates on reliability of the SOM visualization display. New measures are developed for estimating the relative reliability of different parts of the visualization. Furthermore, methods for assessing the reliability of groups of samples manually extracted from a visualization display are introduced.

Finally, a new computational method is developed for a specific problem in HERV biology. Activities of individual HERV sequences are estimated from a database of expressed sequence tags using a hidden Markov mixture model. The model is used to analyze the activity patterns of HERVs.

ABSTRAKTI

Oja, M. (2007): **Eksploratiivisia menetelmiä genomitiedon analysointiin—sovelluskohteena ihmisen endogeeniset retrovirukset.** Väitöskirja, Teknillinen korkeakoulu, Dissertations in Computer and Information Science, Raportti D23, Espoo, Suomi.

Avainsanat: bioinformatiikka, eksploratiivinen data-analyysi, geenieksspressio, ihmisen endogeeninen retrovirus, informaation visualisointi, itseorganisoituva kartta, luotettavuus, oppiva metriikka, piilo-Markov-malli.

Väitöskirjassa on kehitetty eksploratiivisia data-analyysimenetelmiä genomiaineistojen analysointiin, keskittyen erityisesti ihmisen endogeenisiin retrovirussekvensseihin ja geenieksspressioaineistoihin. Ihmisen endogeeniset retrovirukset (human endogenous retrovirus, HERV) ovat muinaisten retrovirusinfektioiden jäänteitä ja ovat nyt osa ihmisen genomia. HERV:eistä tiedetään kovin vähän, mutta niille on löytynyt yhteyksiä joihinkin sairauksiin. Tämä työ tarjoaa menetelmiä HERV:ien ominaisuuksien ja aktivoitumisen tutkimiseen.

Nykyään genomiaineistot ovat niin suuria, että tarvitaan kehittyneitä data-analyysimenetelmiä datan mielenkiintoisten rakenteiden löytämiseksi. Eksploratiivisten menetelmien tehtävä on auttaa luomaan hypoteeseja datan ominaisuuksista. Esimerkiksi ryhmittelemällä geenit samoin käyttäytyvien, ja oletettavasti saman funktion omaavien, geenien ryhmiin voidaan ehdottaa funktio toiminnaltaan ennestään tuntemattomalle geenille. Eksploratiivisen data-analyysin avulla muodostetut hypoteesit voidaan myöhemmin varmistaa yksityiskohtaisempien kokeiden avulla. Sen sijaan yksityiskohtainen analyysi olisi liian hidasta ja kallista suorittaa kaikille geeneille.

Työssä esitetään itseorganisoituvaan karttaan (self-organizing map, SOM) pohjautuvia eksploratiivisia data-analyysimenetelmiä geenieksspressioprofiilien ja ihmisen endogeenisten retrovirussekvenssien visualisointiin ja ryhmittelyyn. SOM-pohjaista lähestymistapaa täydennetään karttavisualisoinnin luotettavuutta arvioivin menetelmin. Uusia mittareita on kehitetty visualisoinnin eri osien suhteellisen luotettavuuden arviointiin. Lisäksi työssä on esitetty menetelmiä, joiden avulla voidaan arvioida käsin kartalta eroteltujen ryhmien luotettavuutta.

Työssä on kehitetty uusi laskennallinen menetelmä tietyn HERV:ien biologiaan liittyvän ongelman ratkaisemiseksi. Yksittäisten HERV-sekvenssien aktiivisuustasot pystytään menetelmän avulla estimoimaan ekspressoituneita sekvenssejä listaavista tietokannoista. Uusi menetelmä pohjautuu piilo-Markov-sekoitemalleihin. Työssä sitä käytetään HERV:ien ekspressioprofiilien estimoimisessa ja analysoimisessa.

Contents

Preface

This work has been carried out in the Neural Networks Research Centre / Adaptive Informatics Research Centre of the Laboratory of Computer and Information Science, Helsinki University of Technology and in the Department of Computer Science, University of Helsinki. The main sources of funding have been the Graduate School in Computational Biology, Bioinformatics, and Biometry (ComBi) and research programmes of the Academy of Finland, in particular the Microbes and Man (MICMAN) and Life 2000. The corresponding project numbers are listed in the acknowledgments of the individual publications. Additional funding for travel from the Nordic Bioinformatics Network, the International Society for Computational Biology (ISCB) and University of Helsinki is also gratefully acknowledged. Of special encouragement have been the personal grants received from Tekniikan edistämissäätiö and Finnish cultural foundation.

Special thanks go to my instructor and supervisor Professor Samuel Kaski, who has taught me everything I know about scientific work. His contribution to my work is evident from the publications in this thesis and his comments during the preparation of the manuscript are highly appreciated.

My sincere compliments belong to the reviewers of the thesis: Professor Juho Rousu and Docent Tero Aittokallio. Their efforts have helped me to improve it significantly.

I am indebted to all my co-authors. Firstly, I am honored to have had a possibility to work with Academician Teuvo Kohonen. On the methodological side, I wish to thank Janne Nikkilä, Jaakko Peltonen, Jarkko Venna, Panu Somervuo, and Samuel Kaski for their help with the theoretical and algorithmic aspects of my work. Special thanks go to Janne and Jaakko for advice and insightful discussions during different stages of the thesis project. On the biological side, Eero Castrén, Garry Wong, and Petri Törönen introduced the world of biological research to me and Jonas Blomberg and Göran Sperber taught me a lot about the fascinating world of retroviruses.

I wish to thank the whole MI research group, the rest of the laboratory and my colleagues at the University of Helsinki for creating an easy atmosphere to work in.

I am grateful to my parents and my sister for their support and encouragement. I also wish to thank all my friends for the refreshing moments and occasions between the work days. Lastly and most of all, I wish to thank my husband Nuutti.

LIST OF PUBLICATIONS

The thesis consists of an introduction and the following publications:

1. Merja Oja, Janne Nikkilä, Petri Törönen, Garry Wong, Eero Castrén, and Samuel Kaski. Exploratory clustering of gene expression profiles of mutated yeast strains. In Wei Zhang and Ilya Shmulevich, editors, *Computational and Statistical Approaches to Genomics*, pages 65–78. Kluwer, Boston, MA, 2002.

2. Samuel Kaski, Janne Nikkilä, Merja Oja, Jarkko Venna, Petri Törönen, and Eero Castrén. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics*, 4:48, 2003.

3. Merja Oja, Panu Somervuo, Samuel Kaski, and Teuvo Kohonen. Clustering of human endogenous retrovirus sequences with median self-organizing map. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03), 11-14 September, Hibikino, Japan*, 2003. Proceedings on CD-ROM.

4. Merja Oja, Göran Sperber, Jonas Blomberg, and Samuel Kaski. Grouping and visualizing human endogenous retroviruses by bootstrapping median self-organizing maps. In *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2004), 7-8 October, San Diego, USA*, pages 95–101, 2004.

5. Merja Oja, Göran O. Sperber, Jonas Blomberg, and Samuel Kaski. Self-organizing map-based discovery and visualization of human endogenous retroviral sequence groups. *International Journal of Neural Systems*, 15(3):163–179, 2005.

6. Merja Oja, Jaakko Peltonen, Jonas Blomberg, and Samuel Kaski. Methods for estimating human endogenous retrovirus activities from EST databases. *BMC Bioinformatics*, 8(Suppl 2):S11, 2007.

7. Merja Oja. *In silico* expression profiles of human endogenous retroviruses. In *Proceedings of the Workshop on Pattern Recognition in Bioinformatics (PRIB 2007)*, volume 4774 of *Lecture Notes in Bioinformatics*, pages 253–263, 2007.

SUMMARY OF PUBLICATIONS AND THE AUTHOR'S CONTRIBUTION

Here a general overview of the author's contributions is given, but the contributions of the other authors are not discussed thoroughly.

In Publication 1 the gene expression profiles of mutated yeast strains are analyzed with a self-organizing map (SOM) based exploratory clustering and information visualization method. The paper demonstrates that the clusters found by the SOM are closely related to the clusters found earlier by hierarchical clustering. The advantage of the SOM is the intuitive visualization of the similarity relationships and cluster structures within the data set. The choice of metric for gene expression data is also discussed. The idea of the research was jointly developed, and the author ran all the simulations and took part in analyzing the results and writing the paper.

Publication 2 proposes tools for addressing two key aspects of data analysis methods for high-dimensional data: i) the visualization performance and ii) the similarity metric. A new trustworthiness measure is introduced and used to compare the performance of visualization methods in displaying gene expression data sets. The learning metrics principle is used to derive a metric from interrelationships among data sets. Yeast genes are visualized with the most trustworthy visualization method, the self-organizing map, in the new metric. The research idea was jointly developed, and the author participated in designing the case studies and conducted the experiments related to the self-organizing map in learning metrics.

Publication 3 shows that the Median SOM, a SOM for sequence data, can be used for studying the mutual relationships of human endogenous retroviruses (HERVs) and their similarities to other DNA elements and retroviruses. The paper demonstrates that a completely data-driven grouping is able to reflect the same kinds of relationships as more traditional biological classifications and phylogenetic taxonomies. The research idea was jointly developed, and the author mainly contributed to its biological aspects. The author took part in designing the experiments and conducted part of the simulations.

In Publication 4 the Median SOM is used to study the relationships of all HERVs and, at the same time, the current HERV classification. The collection of all HERVs is substantially larger and more diverse than the HERV set studied in Publication 3. A novel bootstrap-based method is used to estimate which parts of the SOM visualization are reliable and which are not. The research idea was jointly developed, and the author contributed to all aspects of the work. The author was the main designer of the experiments and conducted all simulations.

In Publication 5 the Median SOM of all HERVs, introduced in publication 4, is studied further. Changes to the current classification of HERVs are recommended and new potential HERV groups are detected. The performance of the SOM is compared to phylogenetic trees. A novel trustworthiness visualization method is introduced and used to estimate which parts of the SOM visualization are reliable and which are not. Furthermore, new measures for estimating the reliability of groups of samples manually extracted from a SOM display are introduced and applied. The research idea was jointly developed; the author contributed to all aspects of the work. The author was the main designer of the experiments and conducted all simulations.

In Publication 6 a generative mixture model, based on Hidden Markov Models, for estimating the activities of the individual HERV sequences from EST (expressed sequence tag) databases is introduced. The performance of the model is validated with experiments on simulated data. The HMM model and a heuristic alternative are used to analyze the activities of real HERVs. The research idea was jointly developed, the author mainly contributed to the initial definition of the HMM structure and the biological formulation of the research problem. The author also suggested biological hypotheses to be studied. The author implemented the method and conducted all simulations and analyzes.

In Publication 7 the Hidden Markov Mixture model developed in publication 6 is extended for the estimation of expression profiles. The extended model is used to study the differential expression of individual HERV sequences of the HML2 group in several tissues. The author was the sole author of the publication.

LIST OF ABBREVIATIONS AND SYMBOLS

List of abbreviations

cDNA	complementary DNA
DNA	Deoxyribonucleic acid
EDA	Exploratory data analysis
EST	Expressed sequence tag
HC	Hierarchical clustering
HERV	Human endogenous retrovirus
HMM	Hidden Markov model
LINE	Long interspersed repeated element
LM	Learning metrics
LTR	Long terminal repeat
MDA	Mixture discriminant analysis
MDS	Multidimensional scaling
mRNA	messenger RNA
MSA	Multiple sequence alignment
PCA	Principal component analysis
PCR	Polymerase chain reaction
PT	Phylogenetic tree
RNA	Ribonucleic acid
RT	Reverse transcriptase
RT-PCR	Reverse transcription polymerase chain reaction
SOM	Self-organizing map
SOM-LM	Self-organizing map in learning metrics
TE	Transposable element
TF	Transcription factor

List of symbols

$\mathbf{x}$, $\mathbf{y}$	data samples
$\mathbf{x}_j$	the jth data sample
$d(\mathbf{x}, \mathbf{y})$	distance between data samples $\mathbf{x}$ and $\mathbf{y}$
$h_{i,j}$	neighborhood function of SOM unit i at unit j
$\mathbf{m}_i$	model vector of ith SOM unit
M_1	the trustworthiness measure
M_2	the continuity measure
M_c	the cluster's consensus measure
N_u	number of samples in unit/cluster u
$d_L(\mathbf{x}, \mathbf{y})$	learning metrics distance between data samples $\mathbf{x}$ and $\mathbf{y}$
M_B	bootstrap-based reliability estimate of a SOM unit
M_T	untrustworthiness value of a SOM unit
$\mathcal{C}$	a group of sequences manually extracted from a SOM display
C_b	compactness of a group of sequences on the bth bootstrap SOM
P_b	purity of a group of sequences on the bth bootstrap SOM
$d(u(i), u(j))$	distance between SOM units $u(i)$ and $u(j)$ along the SOM grid

Chapter 1

Introduction

1.1 General motivation and background

The goal of *molecular biology* is to understand how the cells of a biological organism work. Part of the research is the study of the purpose and *function* of all components of a cell, including proteins and genes. *Functional genomics* is a field of molecular biology that attempts to utilize the vast amounts of data produced by genomic projects to describe functions and interactions of genes and proteins.

The field of molecular biology changed drastically when the new high-throughput techniques, including whole-genome sequencing and cDNA microarrays, were developed in the 90's (Lander, 1999; Butler, 2001; Lockhart and Winzeler, 2000). Currently[1] the sequencing of about 620 species, of which 24 are eukaryotes, have been completed and new species are being sequenced with an ever increasing speed. The genome sequence opens doors to whole new kinds of research directions like the study of gene regulatory elements, human endogenous retroviruses, and other DNA elements outside the gene sequences. The whole-genome sequence also enables the design of microarrays containing all the genes of an organism. Microarrays then allow the simultaneous study of the activity of all the genes.

The high-throughput techniques output huge sets of genomic data. For example, a single human microarray assay, where all human genes are studied in one particular condition, outputs about 12 MB of data. Respectively, the human DNA sequence contains 3 billion base pairs. These data sets are so large that traditional data analysis methods based on simple visualizations and manual browsing of the data are not able to handle them. The amount and dimensionality of data is too large to grasp without the help of computational methods that process and summarize the data and present it to the data analyst in a concise, more easily understandable, format.

This thesis introduces computational methods for analyzing functional genomics data sets. The approaches introduced here can be considered to be forms of *exploratory data analysis (EDA)*. This term was first introduced by Tukey (1977). He defined EDA as a means to explore the data with the help of simple visualizations and summaries; in this thesis the term EDA is used to denote more advanced data analysis methods, including visualization, clustering, probabilistic modeling and combinations of these.

[1] On November 13th 2007, according to NCBI Entrez Genome Project database

EDA enables the analyst to understand the data better and to learn something about the structure of the data, such as clusters of similarly behaving genes. EDA results in the generation of hypotheses about the properties of the data or about the system underlying the data. The analysis result can be, for example, assignment of hypothetical functions to genes in a gene expression study. The hypotheses generated by EDA can later be confirmed with more detailed analyses.

The explorative phase of data analysis is particularly important when the data to be studied is huge, like it usually is in functional genomics, and when the knowledge of the system, where the data is coming from, is limited. Using EDA, the limited resources (money, time, and effort) can be focused on a smaller, well selected set of interesting objects, for example, on a set of genes that were found to be related to a particular cancer subtype in an exploratory analysis of a gene expression data set. The set of interesting genes can be studied more closely in the laboratory using experimental techniques that would be too expensive and time consuming to be applied to all the genes.

In this thesis the focus is on two kinds of genomic data: gene expression data and human endogenous retrovirus sequences. Both data types are described in more detail in this thesis. Here the general properties and challenges posed by these data sources are briefly outlined.

Gene expression data comes from microarray measurements of gene activity. The data is usually measured for the purpose of studying the function of genes. Finding their function is interesting *per se*, but also as a link in the research aiming at new drugs for diseases. Analysis of gene expression data is challenging because the data sets are huge, the data noisy and the system underlying the data, the cell, complex.

Human endogenous retroviruses (HERVs) are virus-derived repetitive elements in the human genome. Little is known about the functions and properties of HERVs. However, as they originate from viruses they might have the potential to harm the human host and have, in fact, been implicated in some diseases including cancer and autoimmune diseases. The amount of HERV sequence data and the length of the sequences pose a challenge to many traditional sequence analysis methods, such as multiple alignments or phylogenetic trees. The HERV sequences, as repetitive elements, are very similar to each other; this causes problems for many laboratory techniques that can be used to study the activation of genetic elements.

This thesis presents computational approaches that can, to some extent, tackle the challenges in analyzing large genomic data sets such as gene expression or HERV data. First of all, a data analysis approach based on self-organizing maps (SOM; Kohonen, 1982, 2001) is proposed for grouping and visualizing gene expression and HERV data.The SOM has previously been used extensively and successfully in large variety of data analysis tasks (Oja et al., 2003). Furthermore, SOMs are well suited for the analysis of huge data sets and able to handle high-dimensional data. The noisiness and high dimensionality of gene expression data is taken into account by choosing the distance measures carefully and by applying the learning metric principle (Kaski and Sinkkonen, 2000, 2004) to automatically learn an appropriate distance measure from data.

The SOM-based exploratory analysis is complemented with estimation of the reliability of visualizations produced by SOM. The reliability of visualization is especially important in EDA, where the visualization is the center point of the analysis process. New measures are developed for estimating the reliability of

different parts of the visualization display. Furthermore, methods for assessing the reliability of groups of samples manually extracted from a SOM visualization are introduced.

In the last part of the thesis a biologically motivated probabilistic model is introduced for estimating the activities of HERVs. The model is able to handle the high sequence similarity of HERVs. The new model can be used to explore the HERVs and make hypotheses about which HERV sequences are active. The HERV expression data obtained by the new model could then be analyzed using the SOM approach introduced (for gene expression data) in the first part of the thesis.

The thesis is on the borderline between biology and computer science in the sense that the biological problems guide the selection and development of the methods proposed here. The SOM-based approaches introduced in this thesis can, however, be applied to other genomic data sources that have similar characteristics. The probabilistic model for HERV sequences introduced in the last part of the work is more purely bioinformatics: the model has been specifically developed for a particular biological problem. Still, it can also be used to study the activities of other repetitive genomic elements besides HERVs.

1.2 Contributions and organization of the thesis

This thesis is about exploratory data analysis approaches for the study of the functions and properties of genes and human endogenous retroviruses. The specific contributions are:

- the application and development of new self-organizing map-based exploratory data analysis approaches for gene expression profiles and HERV sequences
- methods for estimating the reliability of SOM visualization displays and of groups of data samples manually extracted from them
- the development and application of a hidden Markov mixture model for estimating activities of HERV sequences based on databases of expressed sequences (ESTs).

The thesis is organized as follows. In Chapter 2 the biological basis of the work is given. The chapter introduces gene expression data and human endogenous retroviruses. Chapter 3 describes EDA and computational methods that are used or extended in the thesis. Chapter 4 outlines the SOM based exploratory data analysis approach for gene expression data and discusses the choice of metrics and the estimation of reliabilities of different visualization methods. Chapter 5 describes a SOM-based EDA approach for HERV sequence data and introduces new methods for estimation of the reliability of SOM based data analysis. Chapter 6 presents a probabilistic model for the estimation of the activities of HERV sequences from EST data. Chapter 7 summarizes the conclusions of the thesis.

Chapter 2

Genes and human endogenous retroviruses

This chapter gives a brief introduction to the biological application areas studied in this work. The chapter begins by giving the basics of genomes, genes and proteins. Then the measurement techniques and typical features of gene expression data are introduced. Gene expression data is analyzed in Publications 1-2 (see Chapter 4). The rest of the chapter concentrates on human endogenous retroviruses, the main application area in this thesis (Publications 3-7; Chapters 5-6).

2.1 The genome

The genome contains the genetic information of an organism. Genes are a part of the genome. The genome is basically a collection of deoxyribonucleic acid (DNA) molecules called *chromosomes*. A DNA molecule is a sequence of nucleotides that are held together by a backbone. There are four different kinds of nucleotides in DNA: adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T). The sequence formed of these nucleotides contains the genetic information. The DNA molecule almost always exists as a pair of molecules bound together by hydrogen bonds connecting the nucleotides of each *strand* of DNA (Watson and Crick, 1953). The connected nucleotides always pair up A to T and C to G. This *base pairing* property implies a copying mechanism of DNA: a strand of DNA can be built by using the other as template. Ribonucleic acid (RNA) has a structure similar to DNA and uses the same nucleotides, with the exception of using uracil (U) instead of thymine. RNA can base pair with DNA, the mechanism is used in gene transcription where DNA is copied into RNA.

The human genome sequence contains genes, but also other genetic elements such as control elements for genes, telomers, centromeres, and repetitive elements including transposable elements (TE) (IHGSC, 2001). Transposable elements are DNA sequences that contain a machinery enabling them to copy their sequence and insert it to other locations in the DNA. In the human genome the gene sequences actually form only about 2% of the whole DNA sequence whereas the diverse kinds of repetitive elements make up 45% of the sequence (IHGSC, 2001). In this work we will study genes and then human endogenous retroviruses that are one type of TEs.

2.2 From genes to proteins

The process by which a gene's DNA sequence is converted into a functional protein is called *gene expression*. First, the DNA sequence of the gene is *transcribed* into messenger RNA (mRNA), using the base pairing property of DNA and RNA. Then the mRNA sequence in *translated* into protein with the help of ribosomes and transfer RNAs (tRNAs). The tRNAs implement the *genetic code* connecting triplets of RNA (called *codons*) with particular amino acids. The process of transcription and translation is outlined in Figure 2.1.

In a broad sense, a gene consists of a promoter region, a 5' untranslated region (UTR), exons, introns and a 3' UTR (see Fig. 2.2). Only the exons will appear in the final protein product of the gene. The promoter contains binding sites for *transcription factors (TFs)*, the RNA polymerase and other proteins needed for initiation of transcription. The 5' UTR contains signals for translation initiation and the 3' UTR contains the polyadenylation site signaling the end of the gene sequence. Introns are spliced out of the mRNA sequence before translation, leaving only the exons in between the UTRs (see Fig. 2.3). For some genes there exist *alternative splice variants* with different combinations of exons included in the final gene product. An open reading frame (ORF) is a portion of a genome which contains a sequence of nucleotides that could potentially encode a protein.

The genes are basically the same in all the cells of a multicellular organism. However, the cells in different parts (tissues) of the organism are very different, i.e. they exhibit different *phenotypes*. The differences are due to the cell type specific expression of genes: only some of the genes are active in each tissue. Such differential expression of genes is possible because there are mechanisms for controlling (*regulating*) gene expression.

The regulation can happen at any level of the gene expression process. The rate of transcription is controlled by transcription factors. TFs are special proteins that bind the DNA near the gene promoter and interact with each other and the RNA polymerase used to read the gene DNA into RNA. TFs bind to special transcription factor binding sites or to enhancer elements located further away from the gene. There also exist mechanisms that control the gene expression before or after transcription. Chromatin modification processes will reveal the promoter for TFs. After transcription, the rates of mRNA degradation and translation are regulated. Furthermore, the ready protein can still be modified by adding, for example, sugar residues. All these regulation steps affect the final amount and function of the protein.

The gene regulation process is still poorly understood, even though several steps of the process have been identified. The full regulation mechanisms of individual genes are generally unknown. However, more information is available for frequently studied lower organisms, such as the baker's yeast, than for the considerably more complex higher organisms, such as humans.

The information about the functional interactions of genes has been collected to databases (see section 2.5), and is represented there as pathways. A pathway is a cascade or network of genes that affect each other. For example, one gene codes for a transcription factor that then binds to the promoter of another gene and activates it. Examples of cellular processes that have been presented as pathways are signal transduction (response of the cell to external signals) and the cell cycle. Genes that are parts of the same pathway have, by definition, related functions.

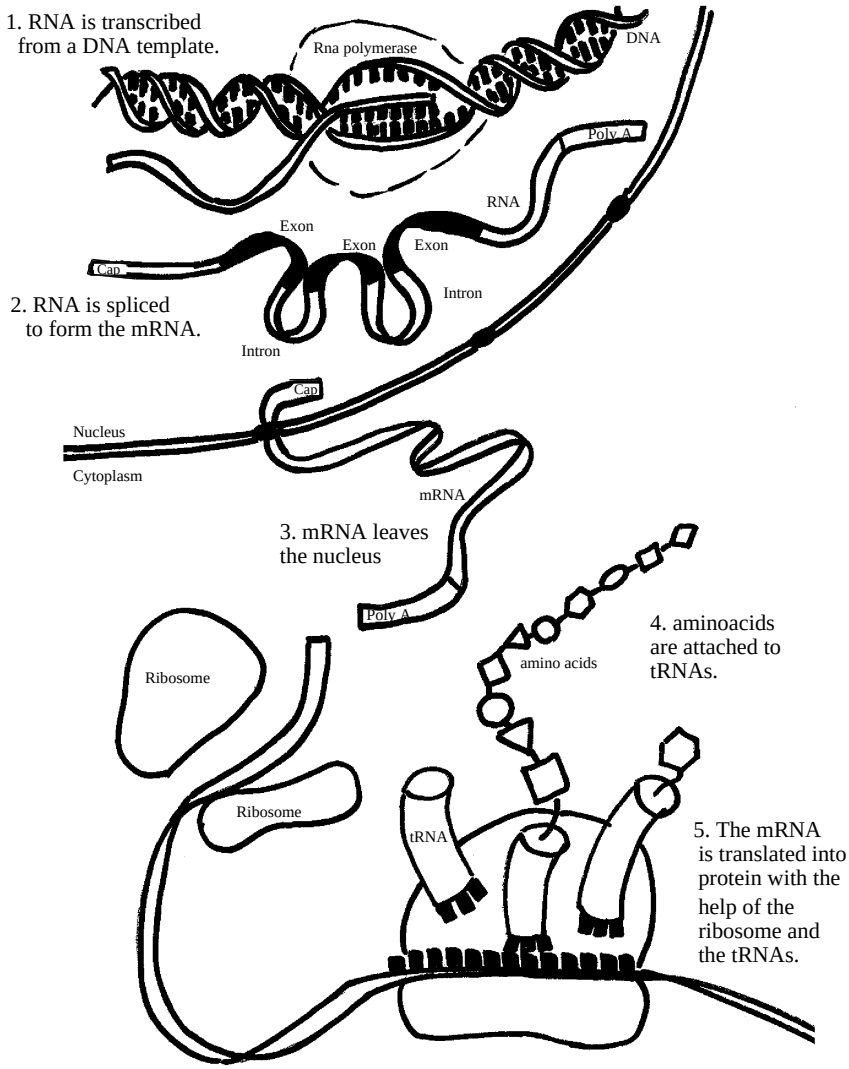


Figure 2.1: Transcription and translation. See text for details.

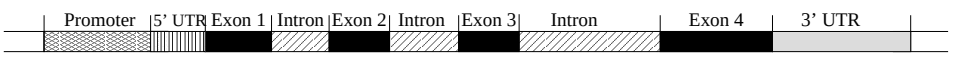


Figure 2.2: Gene sequence in the genome.

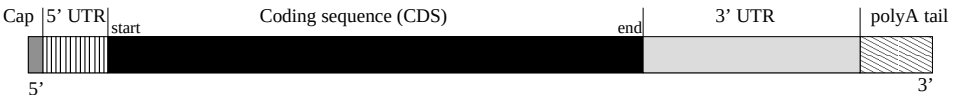


Figure 2.3: Structure of messenger RNA (mRNA). Coloring is the same as in the previous figure.

2.3 Measuring gene expression

Gene expression measurements are usually carried out to study the function of proteins. The idea is that if it is known in which situations the proteins are produced, some clues to their function can be obtained. Furthermore, proteins (genes) that are activated at the same time are likely to be under the same control mechanism and thus parts of the same biological process or pathway. This process of assigning a function to a gene based on co-expression is referred to as the 'guilt-by-association' method (Lockhart and Winzeler, 2000). Actually, the logic goes the other way, i.e., functionally similar genes or genes under the same control mechanism tend to be active at the same time. The 'guilt-by-association' method is a commonly used practice that seems to work, but it should be noted that it is not without fault.

There exist several laboratory methods for the task of determining the *expression levels*, the amount of mRNA, of genes. The expression levels will give indirect evidence about the abundance of the final functional protein product of the gene. It is difficult to measure protein abundance levels on a large scale, but mRNA is easier to study due to the base pairing property of DNA/RNA.

A list of available techniques for measuring gene expression is given for example in Lockhart and Winzeler (2000). In this section only those two methods, microarrays and expressed sequence tags, that are used in this thesis will be reviewed.

2.3.1 Microarrays

The activities of genes can be measured using microarrays, which are able to simultaneously read the expression levels of thousands of genes (Schena et al., 1995; Lockhart et al., 1996; Nature Genetics Supplement, 1999). A microarray provides a way to get a "snapshot" of gene activity. It reports which genes were active, and at what levels, in a cell population at a specific time and in a specific condition.

Microarrays can by manufactured in several different ways, but the most commonly used types of microarrays are cDNA[1] (also referred to as spotted) microarrays (Schena et al., 1995) and oligonucleotide arrays (Lockhart et al., 1996). A cDNA microarray is a small glass or plastic slide upon which a field of spots has been laid. Each spot contains thousands of copies of a *probe*, designed to match the mRNA sequence of a specific gene. The probe cDNA sequence is complementary to the mRNA sequence of a given gene and will base-pair only with that mRNA sequence. In spotted microarrays the probes are synthesized beforehand and are then "spotted" onto the slide. In contrast, in oligonucleotide arrays the probe sequence is synthesized directly onto the slide nucleotide by nucleotide.

The gene expression levels are measured with microarrays using the following procedure (Quackenbush, 2001). First the mRNA is extracted from the sample. The mRNA is converted to cDNA and labeled, typically with a fluorescent dye, before it is hybridized (letting the mRNAs in the sample to base pair with probe sequences on the array) onto the microarray. In the last stage of the process the microarray is scanned with a laser, and intensity of the fluorescence is recorded for each spot. The above applies to both spotted and oligonucleotide arrays with the exception that in spotted arrays two samples are hybridized together onto the

[1] Complementary DNA (cDNA) is constructed by building a strand of DNA using RNA, usually mRNA, as a template

same array. The samples are labeled with different colors and they will compete to bind to the probe sequences. In the end, the intensities of the two colors in a spot will be relative to the amount of that gene's mRNA in the two samples. An example of a scanned cDNA microarray image is shown in Figure 2.4.

The microarray technology is not error free. Typical problems with the arrays include probes that do not match the intended gene and probes that match several genes. Furthermore, dust particles or some other impurities that have ended up on the slide may ruin the hybridization process in some parts of the slide. Imperfections in the measurement process lead to erroneous data values. Some of these can be detected during a quality control process and indicated as missing. Noisiness of microarray data is discussed separately below.

2.3.2 Typical gene expression data

Gene expression is usually measured in several conditions, for example in different tissues, different diseases or patients, different environmental conditions, or at different time points. Nowadays a typical microarray contains all the known genes of an organism. The measurements of gene activity in various conditions form the *expression profile* of that gene. The characteristic profile of a gene is related to its function and can be likened to a fingerprint. Similarly the gene activities in a condition can be seen as a characteristic transcriptional pattern of that condition.

Gene expression measurements output a gene expression matrix with p genes measured in N conditions; a row in the matrix represents the N-dimensional gene expression profile of one gene. The gene expression data matrix is often visualized using green and red to indicate increased and decreased (*up-* and *down-regulated*) expression levels with respect to a reference sample. Such a visualization is show in Figure 2.5.

Gene expression data sets are often huge; p is usually tens of thousands, and N can vary between less than ten and about a thousand. Already the sheer quantity of the data makes it hard to study. For example, when genes are analyzed the number of samples p is prohibitively large for some computational methods. Furthermore, the dimensionality of microarray data poses a challenge for statistical analyses. In a case where the conditions are studied, the data can have a thousand times more dimensions (genes) than samples. This is referred to as the "small N large p problem" (Antoniadis et al., 2003). The self-organizing map based exploratory data analysis approach introduced later in Chapter 4 can handle huge and/or high dimensional data sets. This means that SOMs can be used to analyze either the genes or the conditions of a microarray data set.

Even though the microarray measurement techniques have improved rapidly, the quality of microarray data is often not as high as would be desired. The data contains both biological and measurement noise. Biological noise comes from the natural biological variation between individuals or between different cells. Because the cell is so complex, there are many processes that affect the gene expression levels going on at the same time. Each process will cause variation, some of which is irrelevant from the point of view of the current analysis task. Measurement noise can come from a variety of sources: from techniques used to prepare the biological sample or from the processes of amplification and labeling, or from the array itself. It has been noted that different analysis platforms (cDNA arrays and oligonucleotide arrays from various manufacturers) give different results (Järvinen et al., 2004). Furthermore, microarrays prepared in different laboratories have

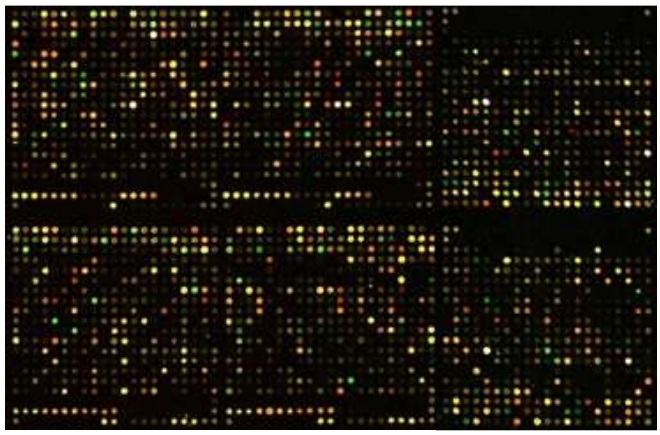


Figure 2.4: A scanned image of a cDNA microarray. Each spot represents one gene, the spots are colored according to the gene expression level. Red means that the gene is more active in the sample than in the control, green marks higher activity in the control and yellow means equal expression in the two cases.

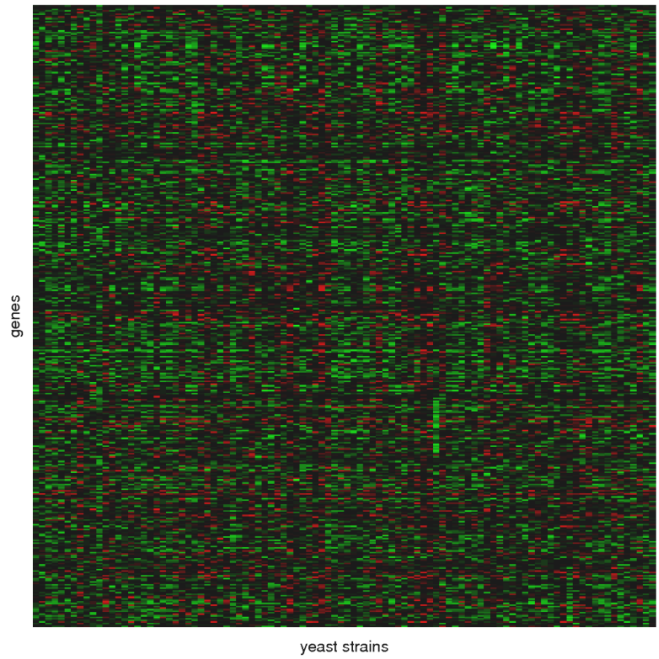


Figure 2.5: Typical gene expression data matrix. Each row represents one gene and each column one array. The size of the matrix is 403 times 98. The data is a subset of the data from Hughes et al. (2000), where gene expression was studied in various mutated yeast strains. The red-green coloring is commonly used to represent microarray data. Red means that the expression log-ratio between sample and control is above 0, green the opposite, and black equal expression.

different characteristics. The between laboratory variation shows up clearly in a visualization of an atlas of gene expression data sets (Venna and Kaski, 2007a). This reflects how much noise biological and measurement variation can bring. The noise is in some cases so high that it drowns the weak signals from lowly expressing genes (Tu et al., 2002). By selecting the metric for comparing gene expression profiles appropriately some of the irrelevant variation can be attenuated. The choice of metric is discussed in Chapter 4.

Preprocessing and normalization issues are very important in microarray data analysis. Normalization is done to make different arrays comparable and to reduce the noise. Quality control steps during the preprocessing ensure that low quality samples are discarded. The thesis does not propose new approaches to preprocessing and normalization of microarray data, even though the issue is briefly discussed in Section 4.2.

2.3.3 Expressed sequence tags

Microarrays cannot be designed if the gene sequences are unknown. In such a case the expressed sequence tagging (EST) technique can be used for gene discovery and gene sequence determination, and also for measuring the activity level of genes.

Expressed sequence tags (ESTs) are short and "noisy" samples of mRNAs present in a cell population. A collection of ESTs is constructed in the following way. First, the mRNA is isolated from the sample and reverse transcribed into cDNA. The cDNAs are inserted into vectors[2] to construct a cDNA library. A random sample of the clones in the library is then selected for one-shot sequencing. The cDNA may be sequenced from either end; forward sequencing produces an EST that matches the beginning of a gene and reverse sequencing products match the end of the gene's mRNA. The EST sequences are relatively low quality (the reverse transcription and/or sequencing steps cause errors) fragments whose length is limited to approximately 500 to 800 nucleotides. Furthermore, the quality of the sequencing usually deteriorates towards the end of the EST. The EST sequences are submitted to **GenBank** and eventually to **dbEST**, a database of ESTs (see section 2.5).

One gene typically produces several mRNA sequences, which leads to several ESTs originating from the same gene. **UniGene** is a database where the ESTs have been grouped to clusters, each representing one potential gene. An EST cluster can be used, for example, to design probes for microarrays (Quackenbush, 2001).

2.4 Human endogenous retroviruses

2.4.1 Retroviruses

Retroviruses are viruses whose genetic material is in the RNA format; for example the human immunodeficiency virus (HIV) is a retrovirus.

Retroviruses are parasites that use the host cell machinery to replicate. The retroviruses hide within the host genome and may lay dormant for a long period before they become infective again. The retroviruses can spread the infection from

[2]In molecular biology a vehicle for transferring genetic material into a cell is called a vector.

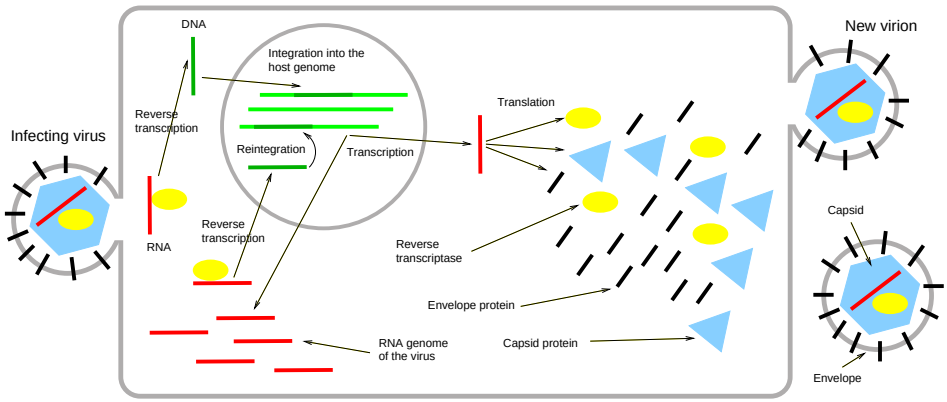


Figure 2.6: Life cycle of a retrovirus. The RNA of the infecting retrovirus is released into the cell, reverse transcribed and then inserted into the host genome. The genes of an active retrovirus are transcribed by the host machinery and retroviral proteins are produced. New virions are compiled from the virus particles.

within the host, i.e., new cells are infected by viruses produced in the already infected cells.

The life cycle of a retrovirus begins when it infects a host by releasing its RNA into the infected cell together with a reverse transcriptase (RT) protein. The RT protein then copies the viral RNA-genome into DNA, which is then inserted into the host DNA sequence. The virus will remain in the host genome and copy itself into other locations. An active virus will use the host transcription machinery to transcribe and then translate its genes into proteins. When enough virus proteins have accumulated new viruses will be assembled from these proteins. Each new virus will also contain a copy of the viral RNA (see Fig. 2.6). When the new virus copies are released from the host cell, it will die.

A typical retrovirus genome consists of long terminal repeat sequences (LTR), one in each end of the genome, and four genes: gag, pro, pol and env (Coffin et al., 1997). The LTR elements function as promoters for the viral genes and also in the process of insertion into the host genome. The gag and env genes code for structural proteins that are used to build new retrovirus particles. The capsid and envelope of the retrovirus are shown in Figure 2.6; the gag gene codes for the capsid proteins and env for the envelope proteins: surface glycoprotein and transmembrane protein. A single retrovirus gene may transcribe several proteins, thus a protease that can cleave amino acid chains to form individual retroviral proteins is needed; this is produced by the pro gene. Finally, the pol gene codes for the reverse transcriptase described in the previous paragraph and for the integrase. Integrase protein is needed in the process of integrating the retrovirus genome into the host DNA. The protease is used to cleave a pol-polyprotein to form the integrase and reverse transcriptase proteins.

Retroviruses have been classified into seven genera: alpharetroviruses (like the avian leucosis virus), betaretroviruses (mouse mammary tumour virus), gammaretroviruses (murine leukemia virus), deltaretroviruses (human T-lymphotropic virus), epsilonretroviruses (walleye dermal sarcoma virus), lentiviruses (human immunodeficiency virus 1) and spumaviruses (human foamy virus).

2.4.2 Human endogenous retroviruses

Endogenous retroviruses are viral sequences within the host DNA (Gifford and Tristem, 2003). In contrast, *exogenous* retroviruses are retroviruses that can exist as virus particles outside the host cells, and are integrated into the host DNA only as a part of their life cycle. An exogenous retrovirus becomes endogenous when it infects a germ line cell and the viral DNA is inherited by the children of the infected host along with the host genome sequence (Gifford and Tristem, 2003).

Human endogenous retroviruses (HERV) are remains of ancient (hundreds or tens of millions of years ago) retrovirus infections. The HERVs, as transposable elements, have been able to move and copy their DNA to other locations in the genome; such copying has yielded several mutated versions of the original virus.

Some of the HERVs have lost the typical retrovirus structure through mutations and various genomic rearrangements, and now contain mutated versions of one or several of the viral genes and zero to two LTRs. The present-day mutated and fragmented HERVs are mainly unable to move and copy themselves. Naturally, older elements have had more time to mutate and generally are less intact. The age of the retroviral element can be estimated from the sequence similarity of its two LTRs: they are identical upon integration and mutate afterwards.

Human endogenous retroviruses have been classified to *HERV groups* (sometimes also referred to as *families*) based on previous studies on their evolutionary origins. A group consists of a set of sequences mutated from the same original infecting virus; a group may contain hundreds of very similar sequences. There are about 40 groups, different sources listing different numbers (Mager and Medstrand, 2003).

HERV groups have been more loosely classified into three main classes. Class I contains endogenous retroviruses that are similar to gammaretroviruses, class II HERVs are similar to betaretroviruses and class III elements are remotely similar to spumaviruses. Class III contains the oldest, most mutated elements, and class II the youngest nearly intact elements.

2.4.3 Function of human endogenous retroviruses

The functions of HERVs can be grouped into two categories: the retroviral function and the genomic function. Retroviral functions of HERVs are related to their origin as viruses. It is highly interesting to know that the human genome contains sequences that can express viral genes in human tissues. The genomic functions are related to their presence in the genome where they can affect the functioning of nearby human genes.

The knowledge about the functions of HERVs is still limited and much work is needed to understand the opportunities and hazards that the HERV may present. The subsections below review what is known about HERV function and what are the most important open questions.

Retroviral functions of HERVs

A retrovirally active HERV could act like a normal infecting retrovirus (Gifford and Tristem, 2003). The active HERV would be able to transpose or to produce virus particles that can re-infect other human cells. Even a partly active HERV can have some of the functions of an active virus, depending on which HERV genes

are expressed. For example, a HERV can transpose if functional RT and integrase proteins are present. Similar HERVs may also be able to work in a co-operative manner, such that the necessary HERV genes are expressed from different genomic locations. Such activity can produce new chimeric HERVs or cause transposition of HERVs that are usually incapable of transposition (Gifford and Tristem, 2003; Blomberg et al., 2005). Active HERVs in the germ line cells can introduce new HERV integrations to the human population, while transposition or re-infection events happening in the somatic cells will only affect the current host.

Retroviral expression has been detected in human tissues, but still little is known about the actual functions of the HERVs or the extent of the expression. As the expression is generally measured for groups of HERVs at a time (see, e.g., Seifarth et al., 2005; Hu et al., 2006), it is not known how many individual active HERVs there are. HERV expression has been studied in various situations. The results show that HERVs are activated in numerous conditions, in both health and disease.

Retroviral expression has been detected in numerous patients suffering from various diseases, for a review see Blomberg et al. (2005) and Nelson et al. (2003). Retroviral transcripts or protein products have been detected in schizophrenia (Shastry, 2002; Frank et al., 2005), cancer (Wang-Johanning et al., 2003; Depil et al., 2002; Patzke et al., 2002; Hu et al., 2006) and autoimmune diseases (Portis, 2002; Christensen et al., 2003). However, HERV expression has been detected also in healthy individuals, see, e.g., Hu et al. (2006). Thus, the connection between HERVs and disease remains highly uncertain. It is not known whether HERVs cause the disease or are expressed because some mechanism normally inactivating HERVs is broken down in diseased cells. In some cases, the expression can be normal in the sense that also healthy cells express HERVs.

The retroviral function of a HERV can also work to benefit the humans. There are at least two known examples of HERV-derived human genes. A HERV env gene, dubbed *syncytin*, has been found to be expressed during an essential step in formation of the placenta (Mi et al., 2000). Another env gene, called *syncytin-2*, is also performing beneficial functions in the placenta (Blaise et al., 2003; Muir et al., 2004).

In order to understand the connection between HERVs and disease, the question that needs to be answered is whether the expression is different between healthy and disease cases. Furthermore, to get insights into potential HERV function it is essential to study how HERV expression differs from tissue to tissue. It would also be informative if the activity could be pin-pointed down to some individual HERV locations. Then also the control of the activation could be studied, starting from the analysis of transcription factor binding sites surrounding the active HERV in the DNA. Basically, HERV gene activation is regulated using the same mechanisms as the "normal" genes.

Genomic functions of HERVs

The HERVs can affect the host by providing alternative promoters and enhancers to host genes, see, e.g., Jordan et al. (2003), Britten (1997) and Medstrand et al. (2005). The HERV LTR sequences, originally designed to activate the virus genes in humans, contain strong activation signals. Thus an LTR sequence nearby a human gene start site may have an effect on the gene activation levels. In some cases the gene is naturally read from the HERV promoter and in some cases the

HERV regulation is abnormal and may cause disease. The regulatory effect of HERVs can be studied by analyzing the genome sequence: the known promoter and enhancer areas around genes can be screened for retrovirus-like sequences as done by Jordan et al. (2003).

In addition to promoters and gene start sites, the HERV sequence can also offer alternative polyadenylation sites to nearby genes, i.e., alternative ending points for a gene in the DNA sequence. In principle, also the splice signals in HERV sequences can be detected by nearby genes causing some portion of HERV DNA to be included into a gene transcript as a new exon (Blomberg et al., 2005).

The HERV sequences offer a pool of sequences for recombination. The HERVs can recombine with similar HERVs in other parts of the genome or with exogenous viruses similar to the endogenous ones (see Bosch and Jobling, 2003; Löwer, 1999). The recombination can result in a more active HERV or it can affect the surrounding genome, for example, by bringing a gene close to a functional retroviral promoter and causing the gene to be more actively transcribed. HERV recombination may also cause a harmful deletion of genetic material (Bosch and Jobling, 2003).

Genomic disruptions happen also when active HERVs transpose to new locations, and hit a gene. A transposition can destroy genes or other necessary elements in its insertion site, thus causing more damage to the cell (Löwer, 1999). Upon insertion into an exon or intron area the retrovirus will knock out the gene or change the protein product. Part of the HERV may even be included into the resulting protein. When inserted near a gene the HERV and especially its LTR elements can have an effect on the regulation of the gene.

2.4.4 Human endogenous retrovirus data

It is estimated that about eight per cent of the human genome is of retroviral origin (IHGSC, 2001). There are a little over 3000 (Sperber et al., 2007) retrovirus integrations that contain at least one retrovirus gene. In addition, there are thousands of single LTR sequences. The exact amount is difficult to estimate, because a conclusive search for single LTRs from the whole genome has not been carried out yet.

The HERV data used in this thesis are HERV sequences detected automatically from the human genome by the program RetroTector (Sperber et al., 2007). RetroTector additionally annotates the HERVs; it estimates the structure of the element (presence and locations of LTRs and viral genes), the age of the element, the intactness of viral gene reading frames etc. It further classifies the HERVs into groups, based on sequence similarity to known representatives of the group. Some sequences remain unclassified because they are not similar to the reference sequence of any group.

The RetroTector program is constructed so that it will find chains of retroviruslike sequences. Each chain is scored and high scoring chains are reported as HERVs. For the data set used in this thesis the cut-off for reporting a chain as a HERV was set rather low. This was done, so that older more fragmented and mutated HERVs would not be missed. In some cases the (low scoring) HERV sequence may contain also other DNA between retroviral segments.

Some challenges of retrovirus data are: the amount of sequences, the length of sequences (full length retrovirus is about 10000kb long), and the similarity of the sequences. Sequences within a HERV group can be almost identical. The

amount of HERV sequence data and the length of the sequences pose a challenge to many traditional sequence analysis methods, such as multiple alignments or phylogenetic inference. In this work an alternative approach is introduced for analyzing large sequence collections (see Chapter 5). The similarity of HERVs causes problems for many laboratory techniques that can be used to study the activation of genetic elements (see the next section). A solution for estimating the activities of individual HERVs is proposed in Chapter 6.

2.4.5 Measuring HERV activity in the laboratory

Several laboratory methods exist for studying HERV activity. The drawback of many of these methods is that they can estimate HERV activity only on the group level, i.e., the activity is a pooled measurement from all members of the HERV group.

Reverse transcription polymerase chain reaction (RT-PCR) is a technique for amplifying a defined piece of a ribonucleic acid (RNA) molecule. The RNA strand is first reverse transcribed into its DNA complement, followed by amplification of the resulting cDNA using polymerase chain reaction (PCR). PCR itself is the process used to amplify specific parts of a DNA molecule. RT-PCR can be used in the task of HERV transcription analysis using primers that will only amplify the desired sequences, for example the members of one HERV group. A primer is a short nucleotide sequence that will base pair with single stranded DNA and allow the synthesis of the second strand to begin. RT-PCR primers are designed such that they will base pair only with the desired sequence and not with other DNA in the sample. In the case of HERVs it is difficult to design primers that would target only one member of the group of highly similar sequences. For this reason the primers for HERVs are designed such that they match all members of the group. Activity of HERV groups has been studied with RT-PCR for example by Hu et al. (2006), Muradrasoli et al. (2006), Andersson et al. (2005) and Forsman et al. (2005).

Seifarth et al. (2003, 2005) have developed a retrovirus chip that is similar to microarrays. The chip is composed of retrovirus-specific synthetic oligonucleotides as capture probes. The retrovirus chip can be used to measure the occurrence of reverse transcriptase (RT)-related transcripts in biological samples of human and mammalian origin. The chip detects expression levels of different HERV groups using probes that target one group specifically.

The presence of antibodies against HERV proteins in the blood serum of a patient shows that the autoimmune system is reacting against HERVs, i.e., HERV proteins are being produced somewhere in the patient. The immune response against HERVs can be studied using the following procedure (see, e.g., Christensen et al., 2003): Short segments of the HERV protein of interest are designed, synthesized and fixed on a membrane. The serum is then added onto the membrane. An antibody in the serum will bind to some of the segments on the membrane and give out fluorescence that can be measured.

All of the above methods target larger groups of HERVs, i.e., give activities for a group of HERV sequences. However, it would be important to get to the level of individual HERV sequences; if it were known which individual HERV loci are activated it would be possible to start to study the control mechanism causing the activation. For example, active sites could share a specific transcription factor binding site not present in the inactive HERVs.

The ESTs can be used to estimate HERV activity similarly as they are used for studying the activities of genes. The basic idea is to count how many EST sequences for each HERV there are in the EST database. However, it is not easy to unambiguously match the EST sequences to the HERV sequences. HERV activity estimation from EST databases is discussed in Chapter 6, and in Publications 6 and 7.

All of the methods described above can also be used to study gene activity and basically many methods used to study the function of genes can be applied to the study of HERVs. After all, the HERVs contain genes. The difficulty is the repetitive nature of HERV which will, of course, have to be taken into account.

2.5 Biological databases

Knowledge about biological systems has been collected into public databases. Sequence information about all human genes and proteins can be found from these resources. There are also databases for gene expression measurement data. But in addition to these there is higher level information available. Some databases list predicted or known functions for genes and proteins. Databases have cross links to each other, connecting sequence data to expression data for example.

Below is a list of databases used in this thesis.

- **Gene Ontology (GO)** project provides a classification hierarchy for genes and proteins. The GO collaborators are developing tree structured ontologies that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. [http://www.geneontology.org]

- **The Kyoto Encyclopedia of Genes and Genomes (KEGG)** (Kanehisa and Goto, 2000; Kanehisa et al., 2004) contains a database of molecular interaction networks such as pathways and complexes. [http://www.genome.ad.jp/kegg/]

- **The MIPS Comprehensive Yeast Genome Database (CYGD)** contains information on the molecular structure and function of the budding yeast *Saccharomyces cerevisiae*. The database offers a functional classification for yeast proteins. The assignment of yeast genes into classes is done manually. [http://mips.gsf.de/genre/proj/yeast/]

- **RepBase** is a database of repetitive elements (Jurka, 2000). It contains consensus sequences for various kinds of repeats, including HERVs. **RepeatMasker** (Smit et al., 1996-2004) is a program that detects repeats from a given DNA sequence. It is based on the RepBase database. Portions of DNA that match some entry of the RepBase well enough are reported as repeats. [http://www.girinst.org]

- **UCSC Genome Browser** is a resource about the human genome. The user can select a genome location and then all kinds of information about the genome area is displayed graphically. Known human genes are shown as well as gene predictions and transcripts that match that area. The human genome has been compared to that of other organisms. The browser shows the amount of conservation between human and chimpanzee, human and mouse, human

and chicken etc. The Genome browser also shows RepeatMasker annotations for that portion of the genome.
[http://genome.ucsc.edu]

- **dbEST** is a database of ESTs. It contains millions of EST sequences for humans alone and millions more for other species.
[http://www.ncbi.nlm.nih.gov/dbEST/]

- **eVOC** is an ontology for ESTs. The eVOC ontologies provide an appropriate set of detailed human terms that describe the sample source of the ESTs.
[http://www.evocontology.org]

Chapter 3

Exploratory data analysis

This chapter lays out the field of exploratory data analysis and introduces the self-organizing maps, the exploratory data analysis (EDA) method applied later in the thesis to group and visualize gene expression data and human endogenous retrovirus sequences. The review of other EDA methods is limited to clustering and visualization methods; they are introduced as comparison methods for self-organizing maps. The methods are described here and used or extended in the subsequent chapters.

3.1 Exploratory data analysis

Exploratory data analysis (EDA) is a principle on how to conduct data analysis. In EDA the data set is visualized and summarized to gain insight into the underlying structure and properties of the data. EDA methods provide a way to look at the data without making restrictive assumptions about the structure of the data space. Furthermore, EDA enables the analyst to draw conclusions about the nature of the process underlying the data set and to generate hypotheses on what kinds of data future experiments studying the process will create. For example, in Chapter 4 hypotheses are made about the function of yeast genes based on a grouping of their gene expression profiles; the hypotheses can later be verified by detailed laboratory experiments.

Examples of low level EDA approaches are simple visualizations such as histograms or box-plots. These visualizations can be used, for example, to unveil the shape of the probability distributions of different components of the data. Simple EDA approaches should be used as a first step in any sort of data analysis task to get an initial feeling of the data and to detect any anomalies in it. Summaries about the means and variances or about the amount of missing values can reveal samples/variables that might best be discarded or normalized.

EDA can also be characterized through its opposites. Historically EDA was considered to be the opposite of hypothesis testing. Hypothesis testing is confirmatory whereas EDA can be seen as the means of *aiding in forming hypotheses* about the properties of the data. EDA can also be seen as the opposite of structured probabilistic modeling, where the model structure is relatively firmly fixed based on prior knowledge. In contrast, probabilistic EDA approaches use very flexible probabilistic models that can learn some structure, like the shape and modality

of the data density, from the data. The flexible models make few assumptions, whereas the structured models try to incorporate all prior knowledge into the system. Of course, probabilistic models are not limited to very flexible and structured models. There is a continuum of models between these two extremes.

3.2 Basics of probabilistic modeling

In probabilistic modeling useful information can be extracted from a data set D by building a good probabilistic model. Machine learning approaches use flexible models M characterized by a set of parameters θ to model the density of the data by $p(D|\theta, M)$. This distribution is often referred to as the likelihood of the data given the model. The task is to learn an appropriate set of parameters such that the model best fits the data D.

The Bayes' theorem (see Gelman et al., 2003) connects the conditional and marginal probability distributions of the data and parameters.

$$P(\theta|D, M) = \frac{p(D|\theta, M)P(\theta|M)}{P(D)}, \tag{3.1}$$

where $p(D|\theta, M)$ is the likelihood, $P(\theta|M)$ is the prior for the model parameters, $P(D)$ is the marginal distribution of the data and $P(\theta|D, M)$ is the posterior distribution of the model parameters given the data D and the model M.

Inference

A common method for inferring the parameters of the model is the maximum likelihood approach where those parameters that maximize the likelihood are selected. For simple cases an analytical solution can be found, but for more complicated cases methods such as the Expectation Maximization (EM) (Dempster et al., 1977) algorithm or sampling approaches must be used.

Bayesian estimation focuses on finding parameters of the posterior distribution of the model. This is usually done by sampling the posterior distribution using MCMC methods (see Gelman et al., 2003). If a single set of parameter values is needed, the values maximizing the posterior probability of the model, the maximum a posteriori estimate, can be used (see Gelman et al., 2003).

Mixture models

A mixture model is a probability distribution that is a combination of other probability distributions (see McLachlan and Basford, 1988; McLachlan and Peel, 2000). The probability distribution of the data $p(\mathbf{x})$, where $\mathbf{x}$ is a sample from the distribution, has the form

$$p(\mathbf{x}) = \sum_{i=1}^{K} a_i p(\mathbf{x}|\phi_i), \tag{3.2}$$

in the case when all K mixture components are from a parametric family of density distributions. Above a_i are the mixture weights that sum up to one and $p(\mathbf{x}, \phi_i)$ are the component distributions, each with its own parameters ϕ_i. Mixture models are used, for example, in clustering and density estimation tasks.

Density estimation

Density estimation is the construction of an estimate of a probability density function of the data (see Silverman, 1986). The unobservable density function is the density according to which a large population is distributed; the observed data is usually thought to be a random sample from that population.

There are parametric and unparametric approaches to density estimation. The popular unparametric Parzen estimate of the density function is constructed by setting an identical kernel (a density function) at the location of each observed data point and by constructing the density function as a mixture of these (Parzen, 1962). The Gaussian distribution is commonly used as the kernel, but other choices are possible as well. In parametric approaches the density distribution is presented using parametric density distributions, usually a mixture of them, see Eq. 3.2 and Silverman (1986).

3.3 Selected exploratory data analysis methods

This section introduces some commonly used exploratory data analysis methods for gene expression data (metric vectors). Most of the methods can be adapted also to work for biological sequence data (or a pairwise distance matrix of the sequences). The review is limited to clustering and visualization methods. They are included as comparison methods to self-organizing maps, that are used in later parts of this thesis to group and visualize gene expression data and human endogenous retrovirus sequences. The basic version of SOM is introduced here, while the adapted versions are introduced later when they are applied in the biological problem setting.

3.3.1 Clustering methods

Clustering is the process of partitioning a data set into groups so that the data samples in one group are similar to each other and are as different as possible from samples in other groups (see Jain and Dubes, 1988). Clustering is a commonly used approach in gene expression analysis. The aim is to find natural clusters of similar genes. The clusters can then be analyzed to discover the function of uncharacterized genes. The assumption is that genes that have similar expression profiles also have similar functions (Lockhart and Winzeler, 2000; Quackenbush, 2001). The gene expression data matrix can also be analyzed in the other direction; the conditions (on the columns of the matrix) are studied instead of genes (on the rows of the matrix). For example, patients are clustered to groups sharing similar gene expression patterns (Bhattacharjee et al., 2001; Golub et al., 1999).

In clustering the number of data samples is reduced by grouping similar samples together, making the manual data analysis task easier: only a small number of cluster representatives need to be analyzed instead of thousands of original data samples.

Hierarchical clustering

There are two types of *hierarchical clustering (HC)* methods: agglomerative, which merge clusters into bigger ones, and divisive, which divide clusters. Of these agglomerative clustering approaches are more common and are also very popular in

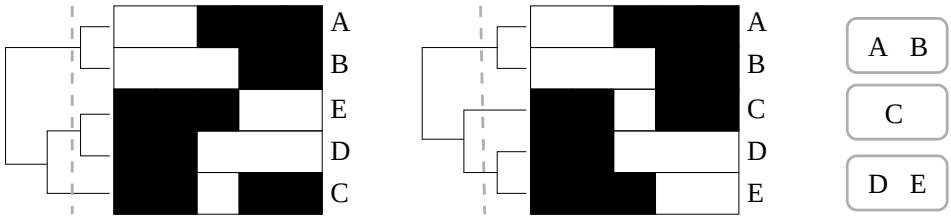


Figure 3.1: Hierarchical clustering result. The figure shows the same hierarchical clustering result twice (left and middle). On the right are the three clusters that were obtained by cutting the dendrogram at the level indicated by the dashed line. The 5-dimensional binary samples are visualized with a black and white matrix that is ordered according to the dendrogram from the clustering. On the left is an ordering that you might get from a standard implementation of hierarchical clustering, the branches and leaves are in no particular order. In the middle the leaves are ordered such that neighboring samples are as similar to each other as possible. Notice how the branches of the tree have been flipped over to achieve the new ordering.

gene expression data analysis (Eisen et al., 1998; Hughes et al., 2000; Bhattacharjee et al., 2001).

In the beginning of a agglomerative hierarchical clustering each data sample forms its own cluster. Then the two closest clusters are merged together to form a new cluster. This process is repeated until all samples belong to one cluster. The decisions on which two clusters or samples to combine depend on the choice of distance measure. An example is the complete linkage method where the distance between clusters is the maximum distance two samples, one from each cluster, can have. The sequence of mergings can be displayed graphically in a tree-like form called the *dendrogram*. The dendrogram can be cut at some level to extract a set of disjoint clusters (see Fig. 3.1 right; Jain and Dubes, 1988; Hand et al., 2001).

The data samples are visualized with the dendrogram as the leaves of the clustering tree (see Fig. 3.1). The branches of the tree are side by side in the visualization, and the order in the branches and leaves induces a linear ordering for the data samples (in the left part of Fig. 3.1 the order of the samples is A, B, E, D, C). The problem is that adjacent samples in the linear order are very dissimilar at those points when leaves from different branches of the tree are placed side by side (for example samples B and E in the left part of Fig. 3.1). Furthermore, the order is not unique, as any branch can be flipped over (like is done for the lower branches in Fig. 3.1: order of the samples has changed from EDC to CDE). These problems can be alleviated in part by optimizing the order of the leaves such that the dissimilarities between neighboring points are minimized (Bar-Joseph et al., 2001, 2003). A result of such an optimization is shown in the middle of Figure 3.1.

When the hierarchical clustering has been performed for both genes and conditions, the dendrogram can be used to order the rows and columns of an expression data matrix as shown in Fig. 3.2.

Another problem with the trees from HC is that they are sensitive to small variations in the data that may affect the agglomeration process, see, e.g. Mangiameli et al. (1996). Furthermore, the dendrograms will be huge for large data sets. It is then hard to extract the essential cluster structures from them.

Probabilistic approaches to hierarchical clusterings have also been presented. These have the beneficial ability to reduce the common problems of hierarchical clustering: sensitivity to noise and the propensity to local maxima. Segal and Koller (2002) apply probabilistic abstraction hierarchies (PAH) to gene expression

```
HLFGQALAQDLSQFSYLDTLALQYMNDLLLAAHSETLCHQATQAHQVLLNFLATCGYKVSKPKAQLC
HLFGQZLAQDLSQFSYLDILVLQYVDDLLLATRSETLCHQTTQALLTSSPPVA-----VSKPKAQLC
HLFGQALAQDLSQFSYLDTLVLQYGDDLILATRSETWCHQATQALLNFLATCG---YKVSKQNAQLC
HLFDQALAQDLGHFSSPGTLVLQYVDDLFLATISEASYQQATLDLLKFLANQG---YKVSRSKAQLC
HLFDQALAQDLGHFSSPGTLVLQYVDDLFLATISEASYQQATLDLLKFLANQG---YKVSRSKAQLC
HLFGQALAQDLGHVSSPGTLFLQYLDDLLLATSSEASCQQATVDLLNFLANQG---YKVSRSKAQLC
HLFGQALAQDLGHFSSPGTLVLQYVDDLLLATSSEASCQQATLALLNFLANQG---Y-ASRSKAQLC
HLFGQALPR-LEPILIPGHLSFG-VDDLLLAAHSETLCHQATQALFNFLATCG---YMVSKPKAQLC
***.* *.: *  .   . * :   :**:**: **:  :*:*          .     .*: :****
```

Table 3.1: A segment of a multiple sequence alignment (MSA) of HERV pol proteins. MSA is a sequence alignment of three or more biological sequences. In general, the input sequences are assumed to have an evolutionary relationship. The common portions of the sequences are aligned to reveal blocks of evolutionary conserved positions (marked with '*' below the columns of the alignment). Columns marked with ':' or '.' are also more or less conserved; they contain amino acids with similar properties. A dash in a column denotes a *deletion* in one of the sequences and a letter in a column full of dashes denotes an *insertion* in one of the sequence.

and biological sequence data. Hierarchical mixture models for text collection have been presented, e.g., by Vinokourov and Girolami (2000, 2002) and Toutanova et al. (2001).

Phylogenetic trees

A phylogenetic tree (PT) is not an exploratory method, but a model for evolution of the species. It is introduced here as a comparison method for a SOM of sequence data. However, the heuristic versions of PT inference algorithms resemble hierarchical clustering, and can be considered to be forms of clustering.

A PT is a tree showing the evolutionary relationships among various species or other entities that are believed to have a common ancestor. In a PT, each node with descendants represents the most recent common ancestor of the descendants, with edge lengths sometimes corresponding to time estimates. Neighbor joining (NJ) is a heuristic algorithm that works on a similarity matrix of the sequences obtained from a multiple sequence alignment (MSA) (MSA is explained in Table 3.1). Once the similarity matrix is obtained, the NJ works similarly as hierarchical clustering.

In addition to the heuristic NJ algorithm more rigorous PT inference methods also exist. The maximum parsimony method (Fitch, 1971) tries to reconstruct the ancestor sequence in each node of the tree; trees that require the least number of evolutionary changes are preferred. Maximum likelihood (Felsenstein, 1981) and Bayesian methods are popular; they introduce a probabilistic model for the evolution of sequences and then solve the parameters of the model to get the PT. These rigorous methods are unable to handle massive amounts of sequences, because of the high computational complexity of the algorithms. Furthermore, the multiple alignment step, that is needed in the heuristic NJ algorithm, is also very heavy to compute. A dynamic programming approach that can find the best alignment has a computational complexity of $O(l^k)$, with k sequences of mean length l. There are several heuristic methods for reconstructing the multiple alignment, for example the popular CLUSTAL W algorithm (Thompson et al., 1994) that constructs the MSA by progressively combining smaller alignments based on a guide tree. The guide tree is obtained from a pairwise similarity matrix using the NJ algorithm.

The multiple alignment step makes even the heuristic neighbor joining method very heavy to use when the the number and lengths of the sequences to be aligned are in thousands.

K-means

In *K-means* clustering the data samples are divided into a predefined number, K, of clusters. Each cluster is represented by a *centroid* which is the mean of the samples in that cluster (MacQueen, 1967; Jain and Dubes, 1988). K-means clustering has been, in addition to HC, used extensively in gene expression data analysis, see, e.g., Tavazoie et al. (1999), Beer and Tavazoie (2004) and Vilo et al. (2000).

The K-means algorithm begins by choosing K initial centroids for the K clusters. The algorithm proceeds by repeating the following two steps. One: each data point is assigned to the cluster whose centroid is closest to it. Two: the centroids are recomputed as means of the samples assigned to the cluster. The algorithm ends when the centroids have stabilized. An example K-means clustering is shown in Figure 3.2.

K-means is very sensitive to the set of initial centroids, and may converge to a different clustering when different initialization is used. The clustering result is a local minimum in the search space. Usually the algorithm is run several times with different initializations to find the best local minimum.

A drawback of K-means clustering, besides the local minima problem, is the fact that it produces just a bunch of clusters without describing the relationships between the clusters. Furthermore, the number of clusters, K, is a free parameter that needs to be guessed beforehand or validated rigorously.

Mixture model-based clustering

In mixture model based clustering the density distribution of the data is modeled with a mixture of distributions, see, e.g., Hand et al. (2001) and Eq. 3.2. Each mixture component corresponds to one cluster and is commonly modeled with a Gaussian distribution. Mixture model based clustering is very similar to K-means if K Gaussians with diagonal covariance matrices are used as the mixture distributions. As opposed to K-means the mixture model gives a soft clustering of the samples. Each sample has a probability to belong to each cluster. A hard clustering can be achieved by assigning each sample to the cluster with the highest probability for that sample. Mixture models have been used in clustering gene expression profiles for example by Yeung et al. (2001).

3.3.2 Visualization methods

Visualization is a powerful data analysis tool. Visualization methods aim to present the data to the analyst graphically in such way that the information processing benefits from the strong pattern recognition capabilities of the human brain.

In a visualization the originally multidimensional and complex data is presented on a (usually 2-dimensional) display. For this purpose the information in the data needs to be reduced in a meaningful fashion as it is impossible to present all the high-dimensional information on a 2-d display. The reduction can be done, for

example by projecting the data onto a 2-d subspace of the original data space. So, visualization and dimensionality reduction usually go hand in hand.

An extensive comparison of different visualization methods can be found in Venna (2007) and publications therein (Venna and Kaski, 2006, 2007a). Here only a few basic visualization methods, those used later in the publications, are reviewed.

Simple visualizations

There exist numerous simple visualization methods, such as histograms, bar plots, box plots, pie charts and scatter plots (Tufte, 1983). The simple visualizations are good tools for the initial stage of data analysis. They can be used to explore general properties of the data in order to get a feeling of the data. The visualizations are, of course, also good when the results of the data analysis task need to be illustrated graphically (like done in Publication 6 for example).

Principal component analysis

Principal component analysis (PCA) is a linear dimensionality reduction method. The goal of principal component analysis (PCA) (Hotelling, 1933) is to find linear projections that maximally preserve the variance in the data. The projection directions can be found by solving an eigenvalue problem $\mathbf{Ca} = \lambda\mathbf{a}$, where $\mathbf{C}$ is the covariance matrix of the data, λ an eigenvalue, and $\mathbf{a}$ the corresponding eigenvector. The data is then projected into the space spanned by the eigenvector corresponding to the two or three largest eigenvalues. An example PCA visualization is shown in Figure 3.2.

In general, the aim of linear dimensionality reduction methods is to obtain a lower-dimensional representation of the data. This is achieved by projecting the data linearly onto the low-dimensional space. The projection is selected such that the resulting visualization is useful for the analysis goal. There exist different linear projection methods, each having their own definition of usefulness; for example, in PCA it is assumed that maximum variance directions are the most interesting.

The linear methods are simple to understand and easy to use, but they perform poorly in cases where the interesting variation in the data forms a non-linear manifold in the original high-dimensional space.

Multidimensional scaling

Multidimensional scaling (MDS) attempts to represent the data as points in a small-dimensional space such that pairwise distances of data points are preserved (see Borg and Groenen, 1997). It can be used for constructing a non-linear projection from the high-dimensional data space to a two-dimensional display plane.

There are several variants of multidimensional scaling that differ in the details of the cost function. The cost function of metric MDS (Kruskal, 1964) is

$$E = \sum_{ij} (d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j))^2, \tag{3.3}$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ is the distance in the input space and $d(\mathbf{y}_i, \mathbf{y}_j)$ is the distance in the output space. Usually Euclidean distance is used. The cost function is minimized

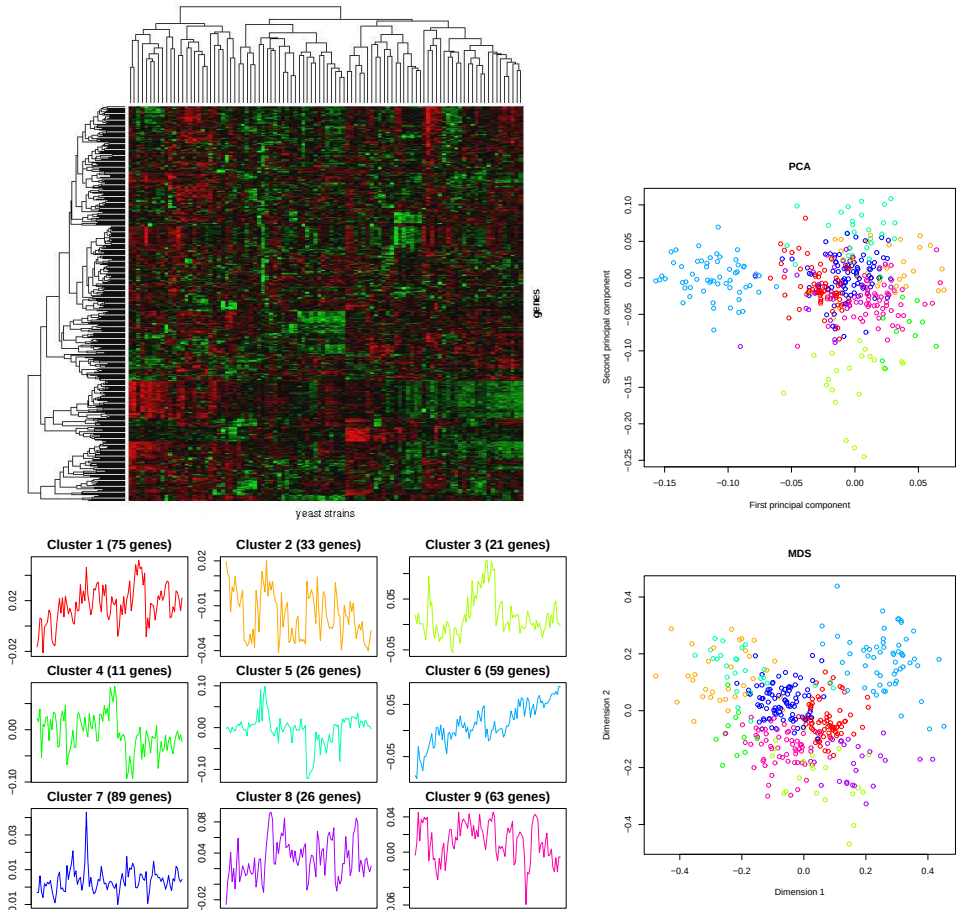


Figure 3.2: Visualizations and clusterings of a gene expression data set. The gene expression data matrix from Figure 2.5 has been explored using hierarchical clustering (top left), principal component analysis (top right), K-means clustering (bottom left) and multi-dimensional scaling (bottom right). For the K-means the gene expression profile of the centroid of each cluster is shown. The last three methods have been applied to the analysis of genes; the coloring for the genes in the figures on the right are derived from the K-means clustering result.

with respect to the $\mathbf{y}_i$:s; they are the representations (locations) of the points $\mathbf{x}_i$ in the output space. An example of a MDS visualization is shown in Figure 3.2.

In Sammon's projection (Sammon Jr., 1969) the mean-square error of the pairwise distances is normalized by the original distances in the cost function. Hence, it emphasizes the preservation of short distances. Non-metric MDS (Kruskal, 1964) attempts to preserve the rank order of the distances. In the cost function of non-metric MDS a monotonically increasing (order-preserving) function f is used. The function acts on the original distances, and always maps them to such values that best preserve the rank order. The cost function then becomes:

$$E = \frac{\sum_{ij}(f(d(\mathbf{x}_i, \mathbf{x}_j)) - f(d(\mathbf{y}_i, \mathbf{y}_j)))^2}{\sum_{ij} d(\mathbf{y}_i, \mathbf{y}_j)^2} \tag{3.4}$$

Above, for any given configuration of the projected points $\mathbf{y}_i$, the function f is

chosen so that the cost function is minimized.

MDS, as a non-linear projection method, is able to find a low-dimensional subspace even from complicated data sets. This is a major advantage over linear dimensionality reduction methods that must limit themselves onto linear subspaces. The visualization with all points presented in a (usually 2-d) display is an intuitive way to show the data. Unfortunately, the coordinate axes of a non-metric MDS display have no meaning and no relation to the original data co-ordinates. This makes it difficult to interpret the observations made based on the MDS visualization.

A disadvantage of MDS and other scatter plot visualizations of the whole data set is the amount of data squeezed onto the display. Even though the dimensionality of data is reduced the amount of points is not. For gene expression data sets this may mean tens of thousands of points on a single display. An ideal solution would be a method that is both able to reduce the dimensionality of the data, visualize it in 2-d and group the data to clusters. The self-organizing map, introduced in the next section is able to do just that.

Manifold learning and other new visualization methods

Manifold learning methods are based on the assumption that the data lies on a low-dimensional manifold in the high-dimensional input space. The goal of manifold learning is to find and unfold this manifold. Local linear embedding (LLE) (Roweis and Saul, 2000) and Laplacian eigenmap (Belkin and Niyogi, 2002) algorithms are examples of manifold learning methods. Both form a presentation of the manifold with the help of local neighborhoods, in LLE each point is represented as a weighted sum of k neighboring points and in Laplacian eigenmap a k-nearest-neighbor graph is formed. When the manifold learning methods are used for visualization the output dimension needs to be set to two (or three). If the manifold actually is of higher dimensionality, the methods may end up in problems (Venna and Kaski, 2007a).

Stochastic neighbor embedding (SNE) (Hinton and Roweis, 2002) is very similar to MDS, but instead of trying to preserve pairwise distances of samples it preserves the probabilities of points being neighbors. Neighbor retrieval visualizer (NeRV) (Venna and Kaski, 2007b) extends the SNE by adding another term into the cost function. In addition to trying to maximize smoothed recall (preserve probabilities of points being neighbors) also smoothed precision is taken into account. The trade-off between recall and precision is familiar from the information retrieval field. Local MDS (Venna and Kaski, 2006) can be considered to be a faster heuristic version of NeRV.

3.3.3 The self-organizing map

The self-organizing map (SOM) (Kohonen, 1982, 2001) is an algorithm that maps high-dimensional data non-linearly onto a low-dimensional map lattice that can be visualized. The SOM can be used as both a non-linear projection and a clustering method; clusters can be extracted from the SOM display either automatically (Vesanto and Alhoniemi, 2000) or manually (see Chapter 5 and Publication 5).

The SOMs have been used widely in various application areas, including speech recognition, image analysis, text, biomedical and business applications. For a

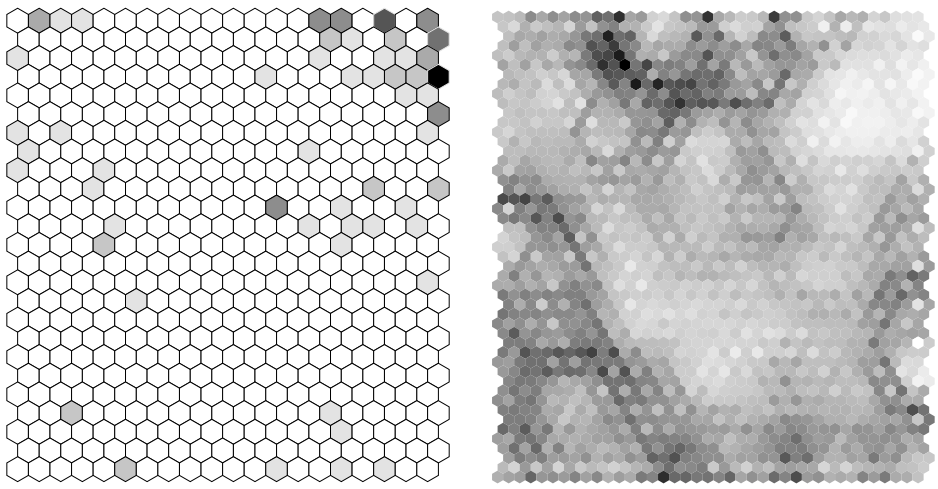


Figure 3.3: The self-organizing map of a gene expression data set. **Left:** A SOM display. Each hexagon presents one map unit. The density distribution of an auxiliary variable is shown on the display. In the darkest map unit, there are 9 genes that belong to the amino acid metabolism class. In white units there are no genes that belong to the class. **Right:** A U-matrix visualization (see section "The SOM visualization display"). Light gray areas display clustered data and darker gray areas are borders between clusters.

comprehensive list of 7718 references to works applying or developing SOMs (see Pöllä et al., 2008; Oja et al., 2003; Kaski et al., 1998).

SOMs for gene expression data are discussed in Chapter 4 and SOMs for biological sequence data in Chapter 5.

The algorithm

The SOM is a discrete lattice of map units (as shown on the left side of Fig. 3.3). There is a *model vector* $\mathbf{m}_i$ attached to each map unit i. A data sample $\mathbf{x}$ is projected onto the SOM display to the map unit having the closest model vector $\mathbf{m}_c$, defined in the basic version of SOM by

$$c(\mathbf{x}) = \arg\min_i d^2(\mathbf{x}, \mathbf{m}_i) \ . \tag{3.5}$$

Here d is the distance measure, which can be for example the Euclidean distance or a correlation based distance.

The input data are represented in an ordered fashion on the map: Map units close-by on the lattice represent more similar samples and units farther away progressively more different samples. The mapping becomes ordered and represents the data after the values for the model vectors have been computed in an iterative training process. In "on-line" type of computation at step t one data sample $\mathbf{x}(t)$ is selected at random, the closest model vector $\mathbf{m}_c$ is found by (3.5), and the model vectors are adapted according to

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) - \frac{1}{2}\, h_{c,i}(t)\, \frac{\partial}{\partial \mathbf{m}_i} d^2(\mathbf{x}(t), \mathbf{m}_i(t)) \ . \tag{3.6}$$

In the above equations a squared form of the distance is used to make derivation easier. For example, in the case of Euclidean distance measure $d^2(\mathbf{x}, \mathbf{m}) = ||\mathbf{x} -$

$\mathbf{m}||^2$, the update rule becomes

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{c,i}(t)(\mathbf{x}(t) - \mathbf{m}_i).$$

In the above $h_{c,i}(t)$ is the *neighborhood function*, a decreasing function of the distance $d(u(c), u(i))$ of the units c and i on the map grid. A typical choice is a Gaussian function

$$h_{c,i}(t) = \alpha(t) \exp\Big(- \frac{d^2(u(c), u(i))}{2\sigma^2(t)} \Big).$$

where $\sigma(t)$ defines the width and $\alpha(t)$ the height of the kernel. The height corresponds to the learning rate that controls how much the model vectors are updated at each stage. Both the $\sigma(t)$ and $\alpha(t)$ decrease monotonically during the iterative training. For more details, variants, and different methods of computing SOMs (see Kohonen, 2001; Pöllä et al., 2008).

The computational complexity of the SOM training algorithm is $O(N^2)$ (Kohonen et al., 2000) when the size of the SOM lattice is selected to be proportional to the number of data samples N. In a general case the computational complexity of the SOM is $O(U^2)$, where U is the number of map units.

Training a SOM

The SOM can be initialized by giving random initial values for the model vectors (Kohonen, 2001). However, an initialization where the SOM grid is already ordered is preferable as it leads to faster learning. In practice, SOM is often initialized by setting the ordered SOM grid on the 2-dimensional subspace spanned by the two largest principal components of the data. This way the map roughly approximates the data density in the beginning, which speeds up learning.

The SOM is usually trained in two phases. In the first, organization phase both the height and width of the neighborhood function are large and a large number of model vectors are adapted at each step. Model vectors of neighboring units on the SOM lattice all learn from the same input $\mathbf{x}$ which results in local smoothing effect on the model vectors in this neighborhood. In continued learning this smoothing effect will lead to global ordering of the SOM lattice (Kohonen, 2001). In the second fine-tuning phase the neighborhoods will be smaller and the adaptation steps focused on a smaller set of map units. The learning rate is also small which means that only small changes are made. Note that the order in which the samples are presented to the algorithm (the learning sequence) affects the training. The algorithm is usually trained with several learning sequences and the resulting maps are compared. The one with the smallest quantization error $\sum_{j=1}^{N} d(\mathbf{x}_j, \mathbf{m}_{c(\mathbf{x}_j)})$ is eventually selected.

There are several things to consider when constructing a SOM. First of all the dimensionality, topology and size of the SOM lattice need to be selected. For visualization purposes a practical choice is a 2-dimensional lattice of hexagons. Hexagonal lattices are preferred because they do not favor horizontal and vertical directions as much as rectangular arrays (Kohonen, 2001). The size of the map, i.e. the number of map units, is sometimes considered to be the number of clusters in the data. A more reasonable approach is, however, to use a larger number of map units and define clusters as groups of map units. When the aim of the SOM analysis is to do visualization in addition to clustering, then the number of SOM

units should be large. For example, in publications 1-5 the average number of samples per SOM units is less than seven in all of the maps. The SOMs with hundreds of units are very flexible and care should be taken to avoid over-fitting. The final width of the neighborhood function with respect to the size of the SOM controls the flexibility of the SOM lattice. The neighborhood width should be set so that it contains more than one unit in the end of the fine tuning phase. This choice makes the SOM lattice stiffer and over-fitting can be avoided.

The SOM suffers from the same local maximum problem as many other methods, but to a lesser degree (Mangiameli et al., 1996). In practice the SOM will generally find a similar cluster structure in different runs of the algorithm. The same behavior was also observed in experiments conducted in this thesis. The SOM has also been empirically found to be fairly robust to the choice of parameters and to noise in the data set (Kohonen, 2001). The same observation was made in this thesis: the exact choice of neighborhood function height and width in different phases of the SOM algorithm does not seem to affect the results. Naturally, the parameters need to be set in the proper range. Guidelines for this can be found from the book by Kohonen (2001).

In the batch-learning version of the SOM (Kohonen, 2001) all model vectors are updated simultaneously. The closest model vector $\mathbf{m}_{c(\mathbf{x}_j)}$ for each data sample $\mathbf{x}_j$ is sought similarly as above and then the model vectors are updated according to

$$\mathbf{m}_i(t+1) = \frac{\sum_j h_{c(\mathbf{x}_j),i}(t)\mathbf{x}_j}{\sum_j h_{c(\mathbf{x}_j),i}(t)} \tag{3.7}$$

The batch version of the SOM does not have a learning rate parameter and thus has less convergence problems (Kohonen, 2001). It is also faster to optimize than the iterative SOM (Kohonen, 2001).

The SOM visualization display

The SOM is usually displayed by presenting each map unit as a hexagon (see Fig. 3.3). Then various properties of the data can be visualized on the SOM display. For example, text labels can be added on the visualization; the labels describe the contents in that part of the map (for an example see Fig. 5.1 on page 56). Another example is the component planes, where the values of one variable are shown with gray shades on the map display. Annotation data, such as class labels of the samples, can also be visualized on the display. For example, in Figure 3.3 the density of a functional class of the genes in each unit is shown.

The cluster structure of the data can be displayed on the SOM using the U-matrix visualization (Ultsch and Siemon, 1990), where the distances between neighboring units are visualized with gray shading (for an example see Fig. 3.3). An extra hexagon is added between each map unit hexagon and shaded based on the distance between the map units. A cluster is an area of the map where the models of neighboring units are close to each other, that is, the extra hexagons have light shading inside a cluster. Borders between clusters appear as dark edges: at the borders distances between neighboring units are considerably larger. The U-matrix visualization of a SOM can be likened to a topographic geographical map: dense clustered areas of the data space can be thought as hills and sparser areas as valleys between the clusters.

Advantages of SOM-based EDA

The SOM has several advantages over clustering methods presented in the previous sections. First of all, the SOM can, at the same time, visualize both the clusters and the similarities of individual samples. Centroid base clustering methods, such as K-means, give only the clusters. In addition, the number of clusters in the data can be estimated visually from the SOM display and need not be fixed beforehand (the number of map units is preselected to exceed the expected number of clusters). Furthermore, the SOM visualizes the relationships of the clusters on a two-dimensional display (see right side of Fig. 3.3). Distances between clusters are shown as well as the relative sizes and positions of the clusters. Other clustering methods are unable to do this; centroid based clustering methods (K-means or mixture models) output an unordered set of clusters and in the visualizations of the dendrograms from HC the clusters have, basically, only one dimensional relationships (see Fig. 3.1 and the text related to it for a discussion about the order of samples and clusters in a hierarchical clustering tree.) However, some might argue that the dendrogram from HC is easy to understand and contains enough information about the relationships of the clusters. HC and centroid based clustering methods have problems in presenting similarities of samples that fall into different clusters. Naturally, the SOM also suffers from this problem for the samples in the units on the border of the 2-d map.

The SOM has advantages over visualization methods as well. The SOM outperforms linear dimensionality reduction methods because it is able to non-linearly map the data onto the 2-d display. Unlike the projection methods, the SOM reduces the amount of visualized points during the mapping; this can be advantageous in many situations. With SOMs tens of thousands of genes can be visualized with only hundreds of map units. Furthermore, the SOM performs both visualization and clustering at the same time. Naturally, projection methods, such as MDS or PCA, can be combined with clustering. A simple approach where the data is first projected onto a 2-d display and then clustered is not generally feasible, as the 2-dimensional representation may be unable to capture the cluster structure (Yeung and Ruzzo, 2001). However, a clustering result can be visualized in a 2-d visualization display as shown in Figure 3.2.

The SOM has also some limitations. The standard SOM algorithm lacks a proper cost function and there is no probabilistic interpretation for it. These facts make the assessment of the uncertainty of the SOM visualization relatively difficult. In this work new approaches for estimating the reliability of SOM visualizations are presented (see section 5.4 and Publications 4-5).

SOM-based data analysis approaches and applications are presented in Publications 1-5. Visualization ability of SOMs is studied in Publications 2, 4 and 5.

3.4 Hidden Markov models

A hidden Markov model (HMM) is a probabilistic model for sequential data (Baum and Petrie, 1966; Rabiner, 1989). Even though the HMM is not generally considered to be an EDA method, in this thesis it is used in an exploratory fashion in Chapter 6: A mixture of HMMs is used to explore a collection of HERVs and suggest which of them are active. On the other hand, the structure of the component HMMs of the mixture is fixed based on prior biological information, contrary to

the definition of EDA methods as flexible models with few prior assumptions.

The HMM is a generative model that produces data following a Markov process. For example, in a first order Markov process the current state of the process depends only on the previous state. The transition parameters govern how probable it is to move from one state to another. The transition parameters are usually collected into a square matrix T where the entry t_{ij} tells how probable it is to move from state i to state j. In addition to the transition parameters, the probabilities of different initial states are needed; these are represented with the vector $\mathbf{a}$ whose entry a_i is the probability of state i to be the first state in a path (sequence of states) through the HMM. In an HMM the states are hidden from the observer: only the emissions from the states can be observed. Each state has a distinct emission distribution. When the emissions are discrete, for example symbols from a limited alphabet, then emission distributions can be parameterized using a matrix E whose entry e_{io} tells how probable it is to emit symbol o from state i. Formally the probability distribution of a sequence of observations $\mathbf{o} = \{o_1, ..., o_l, ..., o_L\}$ given the HMM and its parameters $\theta = (T, E, \mathbf{a})$ is

$$P(\mathbf{o}|\theta) = \sum_{\pi} P(\mathbf{o}, \pi|\theta) = \sum_{\pi} P(\mathbf{o}|\pi, \theta)P(\pi|\theta), \tag{3.8}$$

where $\pi = \{\pi_1, ..., \pi_l, ...\pi_L\}$ is the unknown sequence of hidden states, i.e., the path through the HMM. In the above

$$P(\mathbf{o}|\pi, \theta) = \prod_{l=1}^{L} e_{\pi_l o_l} \quad \text{and} \quad P(\pi|\theta) = a_{\pi_1} \prod_{l=1}^{L-1} t_{\pi_l \pi_{l+1}}. \tag{3.9}$$

Above the left equation holds if observations are conditionally independent given the state sequence. The right equation follows directly from the definition of a first order Markov process. For continuous observations the emission probabilities e_{io} need to be replaced by a continuous distribution $p(o|i) = p(o|\phi_i)$ with some parameters ϕ_i for each state. Note that the mixture model (Eq. 3.2) is a special case of HMMs when the rows of the transition matrix T are all equal, i.e., the probability to move to a state does not depend on the previous state. If the initial value distribution $\mathbf{a}$ is also the same as all rows of T then

$$\begin{aligned} P(\mathbf{o}|\theta) &= \sum_{\pi} [\prod_{l=1}^{L} p(o_l|\phi_{\pi_l})] a_{\pi_1} [\prod_{l=2}^{L} a_{\pi_l}] \\ &= \sum_{\pi} [\prod_{l=1}^{L} a_{\pi_l} p(o_l|\phi_{\pi_l})] = \prod_{l=1}^{L} \sum_{i} a_i p(o_l|\phi_i), \end{aligned} \tag{3.10}$$

i.e., the model is a mixture model for each observation o_l. In such a case the probability to move to a state corresponds to the mixture weights and the states correspond to the mixture components.

HMMs are used in a wide selection of applications, such as speech recognition (Rabiner, 1989), communications engineering, finance, and bioinformatics.

The profile HMM

The profile HMM (Krogh et al., 1994) is an HMM for biological sequences and has a special sequential structure (see Fig. 3.4). The profile HMMs are generally

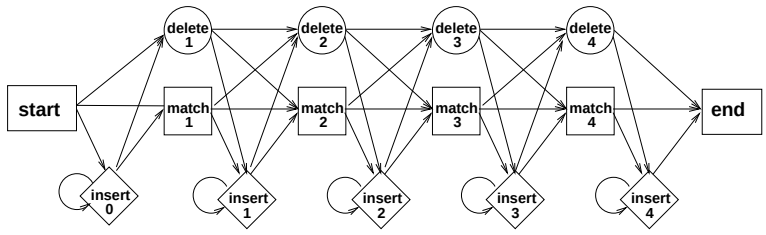


Figure 3.4: Profile hidden Markov model. Match states and insert states are emitting states and emit nucleotides or amino acids according to an emission distribution. Delete states are silent states and do not emit anything.

used to model protein domains[1]. A profile HMM has three kinds of states, match states, insert states and delete states. The delete states are special because they do not emit anything. Note that the transition matrix T of a profile HMM is very sparse as only transitions between nearby states are possible. The profile HMM can, in some cases, be considered to be a probabilistic representation of a multiple sequence alignment (example alignment is shown in Table 3.1): the match states correspond to the conserved amino-acids making up the backbone of the alignment, insert states are used to insert amino-acids not appearing in all sequences and delete states are used to represent sequence positions which are missing from a sequence when compared to the backbone of the alignment.

The Baum-Welch algorithm

The HMM inference task is to learn the model parameters from a set of training sequences that are assumed to have been generated by the HMM model. The HMMs can be trained using the Baum-Welch algorithm, which is practically an Expectation Maximization (EM) algorithm. In the expectation (E) step of the algorithm the expectation over the hidden variables (the path through the model) is computed and in the maximization (M) step the likelihood is maximized with respect to the parameters keeping the expected value of the hidden parameters fixed. The Baum-Welch algorithm is also referred to as the forward-backward algorithm, because in the E-step the likelihood is propagated through the model first from the beginning to the end and then vice versa.

Another inference problem in HMMs is to find the most probable path through the hidden states that could have produced the observed sequence. The Viterbi algorithm handles this task (Viterbi, 1967; Forney, 1973).

HMMs in bioinformatics applications

HMMs are extensively used in bioinformatics applications. Protein domain modeling using profile HMMs was already mentioned above. There are several implementations and databases for the task, for example, SAM (Hughey and Krogh, 1995) is a tool for constructing an HMM model for a set of protein sequences. The Pfam database (Bateman et al., 2002) is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families.

[1] A protein domain is a part of a protein sequence that is conserved in sequence, structure and function in a range of similar proteins.

Other application areas of HMMs in biology include haplotype reconstruction (Sohn and Xing, 2007; Huang et al., 2007), comparative genomics (Lunter, 2007) and models for preprocessing tiling arrays (Shah et al., 2007).

In this thesis mixtures of HMMs are used for the estimation of human endogenous retrovirus activities (see Chapter 5). HMM mixtures have also been used in other bioinformatics applications, for example by Krogh et al. (1994) to cluster protein sequences, and by Schliep et al. (2005) to model gene expression time courses.

The HMMs used in bioinformatics are very diverse. The models most similar to the model used in this thesis are the profile HMM models for proteins domains or families. The main difference is in how the the emission distribution of the match states are defined. In protein domain modeling the match states are learned based on a group of proteins, whereas in the HMM used here the match state emissions are partially fixed based on a given HERV sequence (see Chapter 5). Furthermore, in this thesis the structure of the profile HMM is set based on biological assumptions of the data generation process. The idea of using a mixture of profile HMMs is not new as it has already been used for protein clustering by Krogh et al. (1994).

3.5 Estimating the reliability of exploratory data analysis results

Reliability is a major issue in any data analysis task. In exploratory data analysis it is particularly important to be able to rely on the visualized similarities.

Visualization of similarities of high-dimensional data items, such as gene expression profiles, is a difficult task and usually results in compromises regarding which kinds of relationships to visualize. Reliability of the ensuing visualizations need to be measured so that a visualization technique most suitable for the task at hand can be selected.

Estimation of the reliability of clusters obtained from a clustering method or manually from a visualization is also important. It is necessary to verify that the cluster is truly present in the data and not only an artifact of the clustering method.

In this section measures used to estimate the reliability of visualizations and clusterings are presented. The methods will be discussed more thoroughly in the following chapters where they are applied and extended in the context of exploratory data analysis for genomic data sets.

3.5.1 Trustworthiness and continuity

The compromises made by dimensionality reduction and visualization algorithms result in two kinds of errors: i) samples that were not proximate in the original space are placed close to each other on the display and ii) samples that were proximate in the original data are not close to each other in the visualization. Venna and Kaski (2001) introduced two new measures are to quantify these errors in different parts of a visualization display. The *trustworthiness* measure quantifies the first kinds of error and the *continuity* measure the second kind. It can be argued that the trustworthiness of the visualization is more important of these two in exploratory data analysis.

An area on a display is considered trustworthy if all samples close to each other on the display can be trusted to have been proximate in the original space as well. The trustworthiness of the whole visualization display is measured as a sum of trustworthiness scores over all samples included in the visualization. For each sample $\mathbf{x}_i$ the set of samples $U_k(\mathbf{x}_i)$ that are among the k nearest neighbors of it in the visualization display but not among the k nearest in the original data space is first sought. The unreliability induced by these samples is measured using rank distances: $r(\mathbf{x}_i, \mathbf{x}_j)$ is the rank of the data sample $\mathbf{x}_j$ in the ordering according to the distance from $\mathbf{x}_i$ in the original space. The measure of trustworthiness of the visualization, M_1, is then defined by

$$M_1(k) = 1 - A(k) \sum_{i=1}^{N} \sum_{\mathbf{x}_j \in U_k(\mathbf{x}_i)} (r(\mathbf{x}_i, \mathbf{x}_j) - k) \; , \tag{3.11}$$

where $A(k) = 2/(Nk(2N - 3k - 1))$ scales the values between zero and one.

The projection from a high-dimensional space to a lower-dimensional visualization may have discontinuities. These result in displays where neighborhoods of points in the original space are not preserved. The errors caused by discontinuities are quantified similarly as the errors in trustworthiness. Let $V_k(\mathbf{x}_i)$ be the set of those data samples that are in the neighborhood of the data sample $\mathbf{x}_i$ in the original space but not in the visualization, and let $\hat{r}(\mathbf{x}_i, \mathbf{x}_j)$ be the rank of the data sample $\mathbf{x}_j$ in the ordering according to distance from $\mathbf{x}_i$ in the visualization display. The effects of discontinuities of the projection are quantified by how well the original neighborhoods are preserved, measured by

$$M_2(k) = 1 - A(k) \sum_{i=1}^{N} \sum_{\mathbf{x}_j \in V_k(\mathbf{x}_i)} (\hat{r}(\mathbf{x}_i, \mathbf{x}_j) - k) \; . \tag{3.12}$$

Sometimes the definition of the k nearest neighbors is not unique. There might be ties in the rank ordering, caused by equal distances. In such cases all compatible rank orders are considered to be equally likely. For practical reasons, only the orders that produce the best and worst measures are considered and an average of these is then used. Equal distances occur, for example, in SOM and HC visualizations. When collecting a set of k nearest samples on the SOM display samples are first selected from the same unit. Then the neighboring SOM units are considered in the order of their U-matrix distances (see section "The SOM visualization display" under section 3.3.3).

The trustworthiness and continuity measures were first introduced by Venna and Kaski (2001). Projection methods have been compared earlier for other kinds of data (Mao and Jain, 1995; Goodhill and Sejnowski, 1997). In the earlier comparisons the capability to preserve (all) the actual distances was evaluated, whereas here the criterion is the ability to preserve the proximities (neighborhoods). Furthermore, the previous approaches have not considered the trade-off between the two types of errors made in the projection (the trustworthiness and continuity aspects). The trustworthiness and continuity measures are used for the first time in conjunction with gene expression data in Publication 2. The trustworthiness measure is extended to estimating the reliability of different areas of a visualization display in Publication 5.

3.5.2 Bootstrap

Bootstrap (Efron, 1979; Efron and Tibshirani, 1993) is a resampling method used for estimating the sampling distribution of a given quantity. The bootstrap method is applicable to the following problem: Given a random sample $\mathbf{X} = (X_1, X_2, ..., X_n)$ from an unknown distribution F, estimate the sampling distribution of some prespecified random variable $R(\mathbf{X}, F)$, on the basis of observed data D. The sampling distribution of $R(\mathbf{X}, F)$ is estimated by producing new samples $\mathbf{X}^*$ with replacement from D and computing $R(\mathbf{X}^*, F)$ for each sample. The histogram of $R(\mathbf{X}^*, F)$ values represents the sampling distribution. The sampling distribution can then be used to compute summary statistics for $R(\mathbf{X}, F)$. When bootstrap is applied to clustering, the random variable of interest is the full clustering result.

The bootstrap method has been used in conjunction with clustering (Jain and Moreau, 1987; Levine and Domany, 2001; Bhattacharjee et al., 2001; Monti et al., 2003) to estimate the stability of the discovered clusters. It is assumed that the cluster composition should not change radically between two sets of samples from the same underlying data distribution. Therefore, robustness of a clustering to sampling variability gives support to its validity. However, a new result by Ben-David et al. (2006) argues that stability of a clustering result does not necessarily mean that the clustering would have found the "natural" clusters in the data. The stability may result from an asymmetry of the underlying data distribution, which results in a unique, though incorrect, clustering result. The new result does not, however, invalidate the reasoning that stable results are more likely to be "correct" than unstable results. Thus, it may be sensible to prefer stable clusterings over unstable ones.

Bootstrapping can be used to estimate the stability of both hierarchical and K-means clustering. The idea is to count, for all pairs of samples, how often they appear together (in the same cluster) in the clusterings constructed from resampled data sets. For example, Monti et al. (2003) define a consensus matrix as a square matrix that stores, for each pair of items, the proportion $f_{i,j}$ of clustering runs in which the two items, i and j, are clustered together. The matrix can be used to compute various statistics describing the cluster stability. Monti et al. (2003) define the cluster's consensus as

$$M_c(u) = \frac{1}{N_u(N_u - 1)/2} \sum_{i<j, i,j \in \text{ cluster } u} f_{i,j}, \tag{3.13}$$

where N_u is the number of samples in the cluster u.

The bootstrapping procedure can also be applied to SOMs. In this thesis the bootstrap is used for estimating the reliability of groups of samples extracted from the SOM visualization (see Section 5.4.3) and the reliability of different areas of a visualization (Section 5.4.2).

Chapter 4

SOMs for visualization and clustering of gene expression data

This chapter introduces self-organizing map (SOM) based exploratory data analysis approaches for gene expression data (Publications 1 and 2). The chapter begins by defining the biological problems the SOM-based analysis is trying to address. Then two important aspects of gene expression data analysis, namely preprocessing and the selection of metric, are discussed. Finally, the SOM-based methods are introduced and applied to analyze the functions of yeast genes. The chapter closes with a discussion on the reliability of visualizations produced by SOMs and rival methods, and how the reliability can be measured.

4.1 Problem setting

The function of an organism can be studied with gene expression measurements. Usually this means that the functions of genes are studied by measuring their activities in several conditions.

One typical type of study is a case vs. control study where the aim is to find those genes whose expression is distinctly different in the case samples (such as patients suffering from a particular disease) from the control samples (healthy patients). Genes found to be differentially expressed between the cases and controls are then assigned a disease related function.

Another approach, the one used in this thesis, is to study the genes over larger sets of conditions, for example in different tissues, and then to look at the *gene expression profiles* of the genes, see, e.g., Su et al. (2002). By modeling and comparing the expression profiles it is possible to learn something about the function of the genes. For example, if the fluctuation of the gene expression level over time follows closely the rhythm of the cell cycle we might infer that the gene is working as a part of the cell cycle mechanism. Furthermore, genes behaving similarly in the analyzed conditions might share functions. The reasoning is that genes that are parts of a pathway or respond to a common environmental signal, are likely to be co-regulated. Furthermore, co-regulated genes exhibit co-expression (Quacken-

bush, 2001). Thus, in the spirit of the 'guilt-by-association' method (see Section 2.3) it may be assumed, with reasonable confidence, that a set of co-expressed genes have a similar function in the cell.

In this chapter the self-organizing map is used to cluster genes. At the same time two important issues of exploratory data analysis are addressed, namely the choice of metric and the reliability of visualization displays.

Gene expression data is problematic because the data obtained with microarrays is very noisy and may contain systematic errors. For example, the zero level of gene expression measurements across arrays may be unreliable, making the measurements from different arrays uncomparable. In such a case the distance measure used to compare gene expression profiles should be invariant to the zero level. Another option is to correct the zero level during preprocessing. Good choices of metric for gene expression data are discussed in the following sections.

4.1.1 Compendium of mutated yeast strains

Because many genes have homologous counterparts[1] in different organisms, simple model organisms have been exploited in the analysis of gene function. A popular model organism is baker's yeast, *Saccharomyces cerevisiae*, a unicellular eukaryote species.

One way to study the function of a yeast gene is to remove (or knock out) the gene from the organism and observe the phenotype of the mutant. Giaver et al. (2002) mutated every single gene in yeast and observed the viability and growth of the mutants. One of the problems that has hampered functional analysis is that simple organisms display only a limited number of observable phenotypes and many mutations do not produce any phenotype at all.

Gene expression analysis has been proposed as an alternative means to analyze phenotypes of mutated yeast strains. The global transcriptional response, i.e. the expression level of all the genes, can be regarded as a detailed molecular phenotype of the mutant. Hughes et al. (2000) produced a large compendium of mutated yeast strains, with microarray measurements of the expression of all yeast genes in each mutant. The compendium contains measurements of 300 mutated/treated yeast strains. Of the three hundred strains, 276 were deletion (knock-out) mutants of diverse yeast genes, 11 were mutants with tetracycline-regulatable alleles of essential genes and 13 were strains treated with well-characterized compounds.

Hughes et al. analyzed their data set in an exploratory fashion to make suggestions about gene function. They used hierarchical clustering for this task. In this thesis an exploratory approach based on self-organizing maps is presented.

In Publication 1, the mutated yeast strains were grouped in order to find mutants that induce similar behavior on the rest of the genes. In Publication 2, the gene expression profiles were clustered to groups of genes behaving similarly in the mutant strains. This way also genes that were not knocked out can be studied.

4.2 Preprocessing gene expression data

In this thesis the important issue of gene expression data preprocessing is not considered in depth. Instead, ready-made preprocessing schemes are used for the

[1] Homologous genes have evolved from a common ancestral gene. The homologues have usually also retained the original function of their ancestor.

gene expression data sets analyzed in the Publications. For the yeast compendium data set analyzed in Publications 1 and 2, the approach of the original publication is used.

Generally, the preprocessing of cDNA microarray data (such as the yeast compendium data) has the following steps (Drăghici, 2003): taking the log-ratio of sample vs. control measurements, normalizing arrays so that they have comparable means and scales, averaging the measurements from replicate arrays or probes, and computing standard deviations for the log-ratios. For each of these steps both simple and sophisticated solutions have been proposed.

Preprocessing of yeast compendium data

The yeast compendium data set (described in section 4.1.1) was preprocessed similarly as in the original publication (Hughes et al., 2000). This was done to make the SOM results directly comparable to those of Hughes et al. (2000).

The steps taken in the original publication are briefly outlined below. The natural biological variation of each gene was estimated with replicate measurements of normal yeast strains (untreated normal yeast strain is compared to another untreated strain); this enabled the determination of a gene-specific scaling factor that measures the typical noise level of the gene. The scaling factor could then be used to scale the gene's standard deviation in the measurements of mutated yeast strains; this was done to get more accurate noise estimates. Finally, a p-value was computed to report whether a gene was differentially expressed in the mutant yeast strain. Details about the process can be found from the original publication (Hughes et al., 2000).

In Publication 1 the final yeast compendium data was obtained by first pruning and then normalizing the complete data. First, mutations where only very few genes were differentially (p-value less than 0.01) and/or highly (log-ratio above 3.2) expressed were removed from the data set. Simultaneously, also genes with overall low expression were removed. Then different normalization techniques were tested (for details see Publication 1), and it was concluded that it was a good idea to normalize by the measurement noise, but normalization of gene variance (when the data was analyzed finally in the mutant direction) was not beneficial. The yeast data set used in Publication 2 was preprocessed in a similar fashion.

4.3 Distance measures for gene expression data

In Publication 1 the goal was to cluster mutant yeast strains into groups of (functionally) similar mutants and, at the same time, to maximize the dissimilarity between clusters. The results of such a clustering will depend heavily on the definition of *similarity*.

The most common distance or similarity measures for real-valued data, such as gene expression profiles, are based on either the Euclidean distance or the correlation coefficient. In this section variants of these two measures are considered and their suitability for gene expression data is discussed.

The suitability of different measures depends both on the accuracy of the data and on the goal of the analysis. The questions that need be answered include: (i) Is the zero level of the measurements reliable enough? (ii) Which is interesting, the absolute magnitude of expression ratios or the relative values?

When the focus of the analysis is on the arrays, like when comparing the knock-out mutants, the issue of reliability of the zero level is highly relevant. The scanned microarrays usually have different intensity distributions even though it is reasonable to expect that between two different arrays approximately the same amount of changes in gene expression have occurred. In each array different genes may be up-regulated, but overall the amount of up-regulation is probably approximately the same. The fact that most of the genes are anyway very lowly expressed supports the claim that the means of the arrays should be equal. The differences in the intensity distributions may be the results of, for example, different amounts of mRNA, different settings of the scanner, differences between individual arrays, etc. (Drăghici, 2003). Possible errors in the zero level are usually handled already in the preprocessing step by setting the means of different arrays to be equal. Alternatively, the errors can be taken into account by selecting a metric that is invariant to the zero level.

When the focus is on the analysis of gene expression profiles the reasons for adjusting the zero level are different than in the case of arrays. The zero level is considered uninteresting when the expression pattern of each gene across experiments is more important than the knowledge of whether the gene is primarily up- or down-regulated[2] (Quackenbush, 2001). If the mean expression level of a gene is subtracted from each experimental measurement, the shape of the expression pattern of each gene across experiments is enhanced (see Fig. 4.1 displays A and B).

In Figure 4.1 displays B and C show the same gene expression profile with absolute magnitudes and normalized such that only the relative differences are shown. In this case the shape of the gene expression pattern is very similar even though for one of the genes the fluctuation is stronger. It depends on the analysis task whether the absolute magnitudes are meaningful. It may be reasonable to expect that genes whose expression patterns have similar shapes are under the same control machinery and belong to the same pathway. In that case the shape is the informative feature of the expression pattern, and the magnitude of expression is irrelevant.

Which measure to use?

When the zero level is not very reliable or interesting then the similarity measures should be invariant to it. In the correlation coefficient the average of the measurements is subtracted before the analysis, and the same could in principle be done for the Euclidean measure as well.

If only the relative magnitudes are interesting then it makes sense to normalize the expression profiles to unit length. In the correlation measure this is done anyway, but the normalization can be applied also to inner product and Euclidean measures.

In Publications 1 and 2 three distance measures were selected for comparison: The correlation that is invariant to both the zero level and the scale of gene expressions, the inner product of normalized vectors, which is invariant to scale, and the Euclidean distance (of the original vectors) that takes into account differences in both zero level and scale. The ability to reflect the similarities of functional classifications of genes is used as the criterion.

[2]Up-regulated genes are more highly expressed in the sample (mutant yeast) than in the control (normal yeast). Respectively, down-regulated genes are less expressed in the sample.

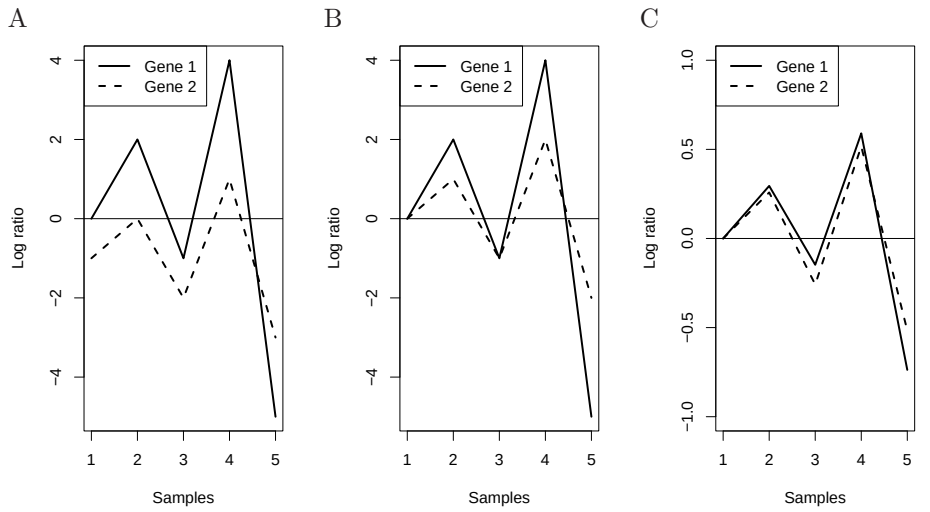


Figure 4.1: **A:** Unnormalized expression profiles. **B:** The zero level of the experiments is considered uninteresting and the mean of both has been normalized to 0. **C:** When only the relative magnitudes are considered interesting the gene expression profiles can be normalized so that the length of the gene expression vectors is one. These profiles have both zero mean and unit length.

Publication 1 discusses the choice of metric for the yeast compendium and arrives at the conclusion that the correlation measure is the most suitable for the analysis of mutant strains (arrays). In the experiments conducted in Publication 2 the correlation measure was found to be the most reliable in the task of comparing gene expression profiles. In both cases the data was ratio-based cDNA microarray data. The work of Hautaniemi et al. (2003) supports the above results; they found the correlation distance to agree better than the Euclidean distance (for ratio-based data) with the similarity relationships assigned by an expert committee of nine biologists. Gibbons and Roth (2002) compared different distance measures in their ability to produce functionally coherent clusters when applied in conjunction with k-means clustering. Contrary to the results reported above it was the Euclidean distance that gave consistently good results for ratio-based data. The correlation distance was found to be best suited for data from oligonucleotide arrays. The difference between the distance measures was, however, not clear for all studied data sets. Furthermore, the performance criterion was different from above. It seems that the suitability of the distance measure is dependent on both the data set and the analysis task.

4.3.1 Learning metrics

The *learning metrics principle* (Kaski and Sinkkonen, 2000, 2004) is a new approach to finding important aspects of data, and expressing them in a way usable for standard data analysis and data mining methods. In general, the learning metrics principle refers to using certain differential-geometric methods for deriving metrics to data spaces, based on the interrelationship between the (primary) data set and auxiliary data. The metrics are called "learning metrics" because they are learned from the two data sets.

Why are learning metrics needed in gene expression data analysis? First of all,

it is known that gene expression measurements in a variety of treatments potentially contain valuable information about the function and co-regulation of genes. The important variation is, however, hidden within all the biological and measurement noise in the high-dimensional expression space. Several processes affecting the gene expression levels are going on in the cell at the same time. Only some of the processes can be studied in one experiment, the rest will cause biological noise. Take for example the cell cycle and changes due to knocking out genes. If only one measurement is taken for each knock-out, it might be that the measurements are at different phases of the cell cycle. In such a setting the cell cycle related variation is noise and only the knock-out induced variance is interesting. A method is needed that is able to discern which variation is meaningful and which is not. For example, when studying the functional similarity of genes in the the knock-out mutation data, the question is which mutations to select, and how to weight the mutations so that the functionally meaningful variation is emphasized and irrelevant variation suppressed. Moreover, the weighting should be different for different genes, that is, at different locations of the expression space. The similarity measures discussed in the previous section are unable to do this, because all dimensions of the gene expression vectors are treated equally.

Using the learning metrics principle it is possible to guide the analysis to directions that will emphasize meaningful variation. For gene expression data the primary data space is the gene expression profiles of the genes and the auxiliary data can be, for example, the functional classifications of the genes; the assumption then is that variation correlating with the functional classifications is meaningful and that all other variation is noise. The learning metrics distance will weight the components of the gene expression vectors locally so that the functionally meaningful variation is emphasized and irrelevant variation suppressed.

Publication 2 was the first to apply the learning metrics principle together with SOMs to gene expression data. The learning metrics had earlier been used to analyze gene expression data in conjunction with clustering (Sinkkonen and Kaski, 2002) and linear discriminant analysis (Kaski and Peltonen, 2003).

The learning metrics distance measure

The learning metrics principle (Kaski and Sinkkonen, 2000, 2004) is based on the assumption that changes in the primary data space are important if they cause changes in another (auxiliary) data space. For example, changes in gene expression (the primary space) are important if they are related to changes in the functional classification of genes (the auxiliary space), see Fig. 4.2.

The learning metrics are defined using the conditional distribution of the auxiliary data given the primary data. The primary data sample is denoted by $\mathbf{x}$ and its functional class by c. During learning, the data occurs in pairs $(\mathbf{x}, c)$. The squared distance measure of the data space is changed locally to measure the important differences, that is, the differences among the distributions of the functional classes $p(c|\mathbf{x})$. When the differences are measured by the Kullback-Leibler divergence D_{KL}, the distances become locally

$$d_L^2(\mathbf{x}, \mathbf{x} + d\mathbf{x}) \equiv D_{KL}(p(c|\mathbf{x})||p(c|\mathbf{x} + d\mathbf{x})) = d\mathbf{x}^T \mathbf{J}(\mathbf{x}) d\mathbf{x} \,, \tag{4.1}$$

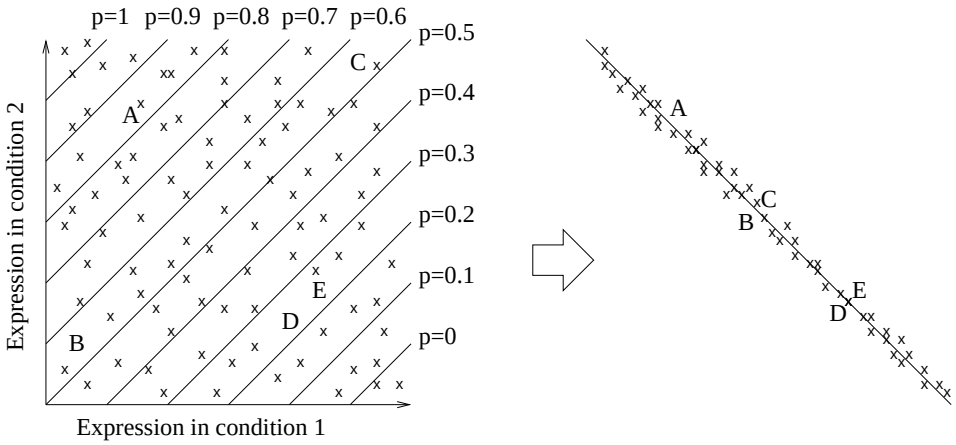


Figure 4.2: Schematic illustration of the change of metric using functional classes of genes (in this case there are only two classes) as the auxiliary data. The expression of five genes, A, B, C, D, and E, has been measured in two treatments. The lines represent the equiprobability lines of the class distribution. In the top left corner of the 2D expression space all genes belong to class 1 and in the bottom right corner all belong to class 2. In the new metric that takes the contour lines into account (on the right) the distances along the direction where the class distribution does not change have been reduced practically to zero. In the new metric the genes B and C are much closer to each other than A and C. The distance of B and C is equal to the distance between D and E.

where $\mathbf{J}(\mathbf{x})$ is the Fisher information matrix

$$\mathbf{J}(\mathbf{x}) = E_{p(c|\mathbf{x})} \left\{ \left(\frac{\partial}{\partial \mathbf{x}} \log p(c|\mathbf{x}) \right) \left(\frac{\partial}{\partial \mathbf{x}} \log p(c|\mathbf{x}) \right)^T \right\} . \tag{4.2}$$

The Fisher information matrix defines the local scaling of the directions of the input space at the point $\mathbf{x}$. For details see Kaski et al. (2001), Kaski and Sinkkonen (2004), Peltonen et al. (2004) and Peltonen (2004). The conditional distribution $p(c|\mathbf{x})$ can be computed using the Bayes rule from a standard estimator of the joint distribution, such as the Mixture Discriminant Analysis (MDA2) (Hastie et al., 1995), or obtained directly from a "mixture of experts" (Peltonen et al., 2002). The metric can in principle be extended to non-local distances by computing path integrals or by approximating the non-local distance over T segments of the direct line connecting two samples (T-point approximation; Peltonen et al., 2004). The complexity of the path integral version is $O(N^3)$ while T-point approximation is T-times heavier than the local approximation.

For computational reasons only the local approximations are used in this thesis, i.e., Eq. 4.1 is used even for large $d\mathbf{x}$. The approximation has worked satisfactorily for nearest-neighbor searches in empirical tests (Peltonen et al., 2004), particularly when complemented with a kind of regularization: in practice the metric will often be singular for very high-dimensional spaces, and hence a portion of the Euclidean distance is added to it,

$$d_L^2(\mathbf{x}, \mathbf{x} + d\mathbf{x}) \equiv d\mathbf{x}^T \left[\lambda \mathbf{I} + (1 - \lambda) \mathbf{J}(\mathbf{x}) \right] d\mathbf{x} , \tag{4.3}$$

where $\mathbf{I}$ is the identity matrix. The coefficient λ is selected using a validation set. The local-approximations suffice well for the SOM algorithm as the close-by distances are the most important for the winner search.

Gene expression data is often normalized such that the length of the gene expression vectors is one (see section 4.3), i.e. the data lies on the surface of a hypersphere. For such data, the density estimators should also be defined on the hypersphere. Technically, instead of using Gaussian kernels, the so-called von Mises-Fisher kernels that are analogs of Gaussians on the hypersphere (Mardia, 1975), can be used (Sinkkonen and Kaski, 2002).

Related approaches

A lot of research has been conducted on methods that extract features, and also some work on learning of metrics. Approaches related to the learning metrics are briefly reviewed here. For a more extensive discussion, see Peltonen (2004) and Sinkkonen (2003).

First of all, different feature extraction and selection approaches are related to learning metrics in the sense that they try to discard irrelevant features. The feature selection methods optimize a transformation into a lower-dimensional output space with a fixed global metric. In contrast, in learning metrics the features are weighted locally. Traditional unsupervised feature extraction methods, including PCA, have been reviewed by Becker and Plumbley (1996). A review on the similarity of feature extraction methods to learning metrics can be found from Peltonen (2004).

Distance metric learning (DML) is an approach similar to learning metrics. In DML a global transformation matrix $\mathbf{A}$ is used to replace $\mathbf{J}(\mathbf{x})$ in Eq. 4.1; various criterions have been proposed for the selection of an optimal $\mathbf{A}$. DML has been applied to clustering with side information in Xing et al. (2003) where the side information is defined in terms of similarity, but the application is to categorized data as in the learning metrics setting. In Schultz and Joachims (2004) a distance metric is learned from relative comparisons. It has also been applied to comparisons derived from categorized data, although some other constraints are added. The DML methods use a global metric whereas the learning metric is a local metric which provides more flexibility.

The genre of genomic data fusion also comes near to the learning metrics principle in the sense that an additional information source is used to derive more informative pairwise-distances. One way to combine similarity information obtained from diverse sources is to present each data source with a kernel and then use a (weighted) sum of the kernels as the final distance presentation, see for example Lanckriet et al. (2004). Kustra and Zagdanski (2006) apply a similar approach: a simple weighted sum of GO annotation derived distances and gene expression distances is used in the task of clustering genes. Shiga et al. (2007) use interaction data as a prior in the process of clustering gene expression data. In the co-clustering method of Hanisch et al. (2002) distance measures are combined using a sum of distances that are first transformed using a sigmoidal function. The transformation emphasizes similarities that are shared between the data sources.

In biclustering both the conditions and the genes of a gene expression data set are clustered simultaneously. The idea is to use only a subset of conditions to describe a cluster of genes. The rationale behind this approach is the observation that a set of genes behaves in a co-ordinated fashion only in some situation (like in one part of the cell cycle) and may have different interaction partners in a different situation. The connection between biclustering and learning metrics is the idea of selecting different features at different locations of the data space (in biclustering

for a cluster at a time). Good reviews of biclustering methods are Madeira and Oliveira (2004) and Tanay et al. (2006).

4.4 SOMs for gene expression data

This thesis introduces a SOM-based data analysis procedure for gene expression analysis. The SOM-based data exploration tools can be used to analyze a genome-wide expression data set and visualize its essential properties.

Compared with most clustering methods the SOM has the advantage that in addition to extracting a set of clusters it also visualizes the relationships between them. The SOM display is an overview of the similarity relationships and cluster structures of the data set. Such visualizations can be used to study clustering of gene expression patterns as done in Publications 1 and 2.

SOMs have been applied to gene expression data previously, for example, by Tamayo et al. (1999); Golub et al. (1999); Nikkilä et al. (2002). The first works Tamayo et al. (1999); Törönen et al. (1999) focused on the ability of the SOM to place similar clusters close to each other, and Golub et al. (1999) used the SOM in a suboptimal fashion: the SOM consisted of 4 units and was used purely as a clustering method in the task of finding subtypes of a cancer. These early works considered each SOM unit as its own cluster. The one unit one cluster approach is also commonly seen in publications that apply SOMs to gene expression data using the GENECLUSTER software (Tamayo et al., 1999), see for example Sesto et al. (2002). In contrast, in this thesis clusters are defined as groups of SOM units. Furthermore, here the focus is on the visualization aspect of the SOM, similarly as in Nikkilä et al. (2002) and Hautaniemi et al. (2003). New in this thesis is the application of SOMs to the visualization of transcriptional patterns of different conditions (mutant yeast strains) (Publication 1), whereas Nikkilä et al. (2002) and Hautaniemi et al. (2003) used the SOM to analyze gene expression profiles. The task of analyzing conditions is more difficult than the analysis of genes, as the vectors representing conditions are usually of considerably higher dimensionality. The combination of SOMs and learning metrics for the analysis of gene expression data is also new (Publication 2).

4.4.1 SOM of yeast knock-outs

The SOM is used to analyze the knock-out mutants in Publication 1. The SOM is computed in the metric that was found to correlate well with the functional classification of genes. The inner product (of vectors of unit length) and correlation distance were found to be preferable to the Euclidean distance for analyzing gene expression (Publication 1). In the comparison correlation distance was the best, but the difference to inner product was not statistically significant. Inner product of vectors of unit length was selected as the distance measure to make it easier to compare the SOM results with those of Hughes et al. (2000) who also used the inner product.

Inner product SOM

The self-organizing map can be implemented in the inner product metric using the following winner selection criterion

$$c(\mathbf{x}) = \arg\min_i \ \mathbf{x}^T \mathbf{m}_i \ . \tag{4.4}$$

and the steepest descent update rule

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{c,i}(t)\mathbf{x}(t) \ . \tag{4.5}$$

In the case where the data lies on the unit sphere, the model vectors will also need to have unit length. Thus the length of $\mathbf{m}_i(t+1)$ is normalized after the update.

Visualization of yeast knock-outs

Figure 4.3 shows a SOM visualization of yeast knock-outs. The black contours are the hierarchical clustering (HC) results from the original publication (Hughes et al., 2000). As can be seen the SOM finds the same structures as the HC. In addition, the SOM visualizes the relationships of the clusters in two dimensions and at the same time represents the similarities of the mutants.

4.4.2 SOM in learning metrics

In this thesis the learning metrics principle is used in conjunction with SOMs to analyze the yeast gene expression profiles from the knock-out measurements. The SOM in learning metrics has previously been used to analyze financial data by Kaski et al. (2001) and diverse data sets by Peltonen et al. (2002).

To use the SOM with learning metrics, it is only necessary to substitute the learning metrics distance to the SOM formulas. The best-matching unit is sought in the new metric by

$$c(\mathbf{x}) = \arg\min_i d_L^2(\mathbf{x}, \mathbf{m}_i) \ . \tag{4.6}$$

and the steepest descent update rule for learning metrics (using local approximations for the distances) turns out (Kaski et al., 2001) to be the same as in the Euclidean metric

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{c,i}(t)(\mathbf{x}(t) - \mathbf{m}_i(t)) \ . \tag{4.7}$$

For gene expression data lying on the unit sphere, the update is applied on the tangent plane, and the results are transformed back to the hypersphere by normalizing the $\mathbf{m}_i(t+1)$ vectors to unit length. It can be shown that the resulting update rule moves $\mathbf{m}_i$ toward $\mathbf{x}$ along the shortest route on the hypersphere, such that their angle reduces by the fraction given by $h_{c,i}(t)$.

SOM-LM of yeast genes

Figure 4.4 displays the yeast genes with SOM in learning metrics (SOM-LM). The visualization reveals several clusters of mutually similar genes. The marked clusters are analyzed more closely in the publication and found to contain functionally coherent groups of genes. The clustering can be used to assign hypothetical functions to unknown genes. The novelty in the display, compared with standard SOM

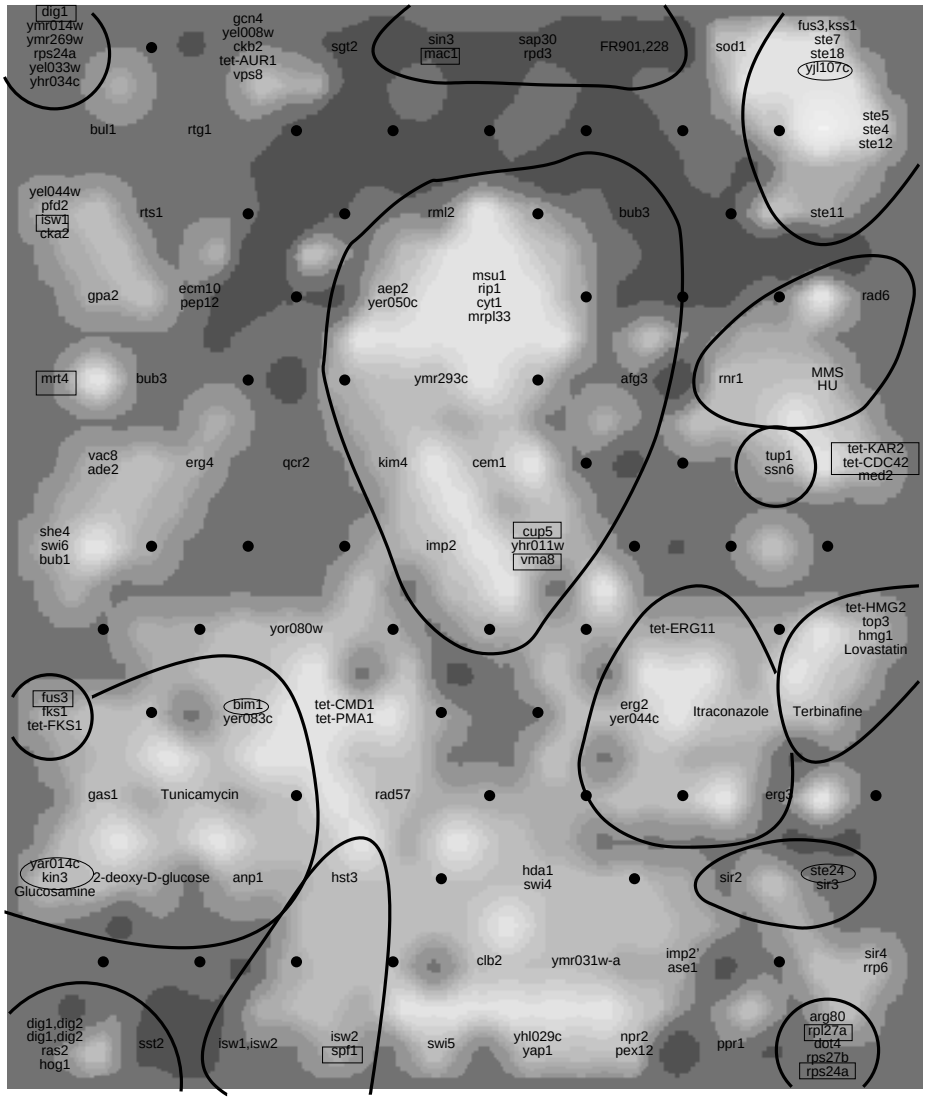


Figure 4.3: A smoothed U-matrix visualization of the SOM of 127 yeast knock-outs. The areas encircled with hand-drawn lines are the clusters reported earlier in the literature (Hughes et al., 2000). The SOM finds the earlier clusters and additionally suggests new groupings (Publication 1). The labels on the map are the names of the genes that have been knocked out in the yeast strain, and the dots are empty SOM map units. White shade denotes high density of the data (clusters) and dark low density (sparse, unclustered area in the data space). The boxes and the circles denote differences between the SOM and the literature clusters: the boxed treatments were grouped to a different cluster in the literature, and encircled treatments are additions to the clusters from the literature. The figure is taken from Publication 1.

displays of gene expression data, is in the metric. Proximate genes both behave similarly in the mutation experiments, and are likely to have similar functional classes.

The SOM-LM was compared to the basic SOM. The ability to predict a class for unknown genes using the SOM display was estimated (see Publication 2). The results show that the new metric yielded more accurate results.

4.5 Reliability of visualizations

Visualizing similarities in high-dimensional (from a few to hundreds of dimensions) data items is a difficult task since the displays can be at most three-dimensional in practice. In particular, it is impossible to project the samples in such a way that all similarity relationships are preserved. Hence, the methods need to make compromises regarding which kinds of relationships to visualize.

On one side of the coin, the visualizations should be trustworthy, in the sense that samples appearing similar (proximate) in the visualization can be trusted to be similar in actuality. The other side of the coin is whether all original proximities become visualized. This dualism is analogous to precision and recall in information retrieval and classification.

The trustworthiness and continuity measures, introduced in section 3.5.1, can measure how well the visualization display performs in preserving the neighborhoods. The measures are computed as a function of k, the number of nearest samples the person looking at the visualization will consider 'proximate'. Results are reported below for several values of k.

4.5.1 SOM is trustworthy

The trustworthiness of the SOM was compared to that of three other visualization methods: Sammon's projection, non-metric MDS, and hierarchical clustering. In all cases the inner product (correlation) similarity measure, that was found to be suitable for gene expression data analysis (see section 4.3), was used. Sammon's mapping and non-metric MDS were selected to represent MDS methods since they have beneficial properties; Sammon's mapping emphasizes the preservation of short distances which are the focus of the trustworthiness measure as well. Non-metric MDS tries to preserve rank orders of distances, which is the error measure used.

The trustworthiness of visualizations of the yeast knock-out data are shown in the left part of Figure 4.5. The trustworthiness of relatively small neighborhoods, of the order of some tens of genes, is the most important, because people looking at displays such as Figure 4.4 will consider these small neighborhoods most saliently proximate. In this range, hierarchical clustering is the best for the smallest neighborhoods ($k < 10$), and SOM after that. The excellent performance of hierarchical clustering at very small neighborhood sizes was to be expected as it explicitly connects the closest points first. In hierarchical clustering two definitions for distances in the visualization were used, the ultrametric distance that takes the tree into account (for details see Publication 5) and a linear ordering along the leaves. The linear order is not unique as was pointed out in section 3.3.1. Here the leaf order was fixed using a method recommended by Eisen[3]: in non-unique cases the order provided by a one-dimensional SOM is used.

[3]See the documentation of the program package at http://rana.lbl.gov/

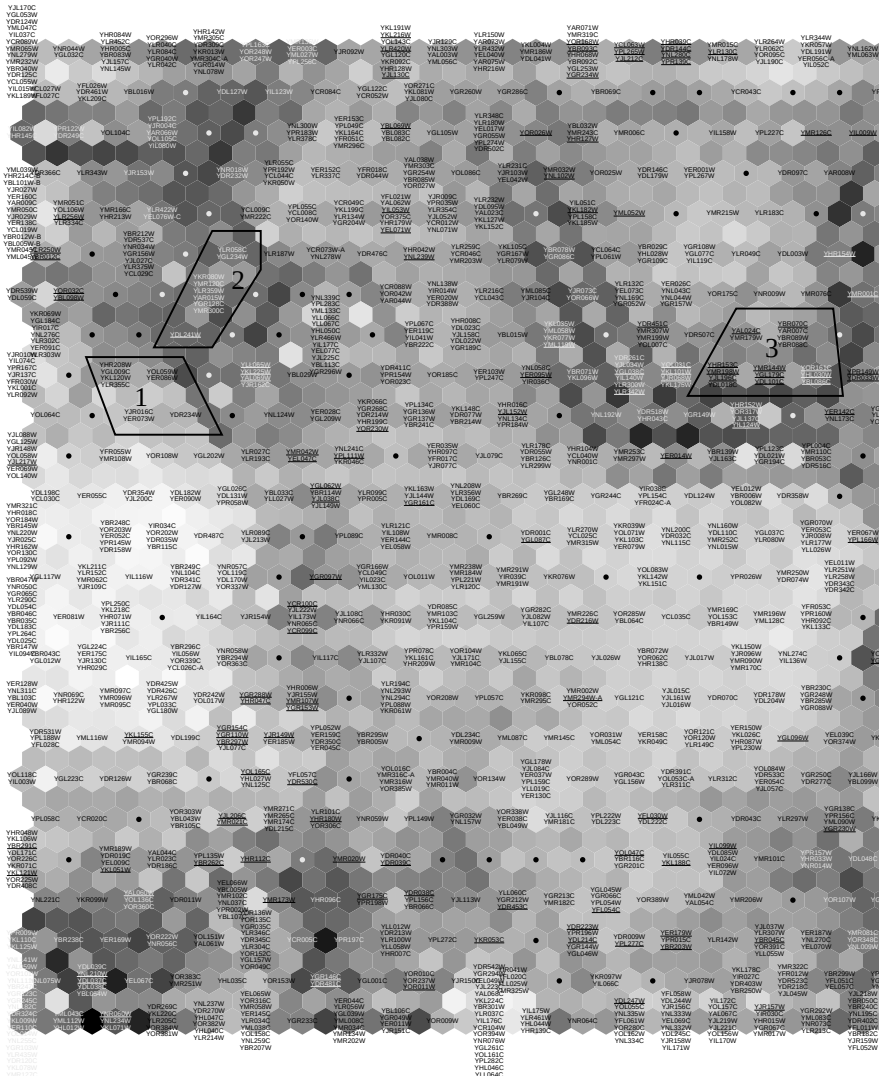


Figure 4.4: U-matrix visualization of the SOM-LM of gene expression data measured from 300 knock-out mutations. The functional classification of genes is used as the auxiliary data in the learning metrics distance. The underlined genes are the ones for which the metric changed the most in comparison to the inner product one. The enumerated clusters are sample clusters: 1: A cluster associated with mitochondria, 2: Localization of purine biosynthesis pathway, and 3: An area where the metric has changed. Most of the genes in area 3 have an unknown function; some are associated to transcription and DNA repair. The figure is taken from Publication 2.

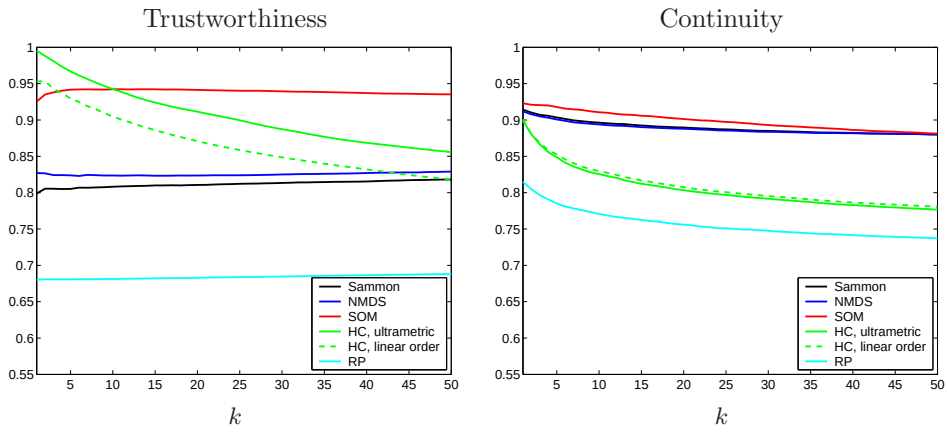


Figure 4.5: **Left:** Trustworthiness of the visualized similarities (neighborhoods of k nearest samples). **Right:** Capability of the visualizations to preserve the similarities (the neighborhoods of size k) of the original data space. Sammon: Sammon's mapping, NMDS: non-metric multidimensional scaling, SOM: self-organizing map, HC: hierarchical clustering, with the ultrametric distance measure and with the linear distance measure. RP: Random projection, i.e. projection of the samples to a randomly selected 2-dimensional plane. RP is the approximate worst possible practical result (the small standard deviation over different projections, approximately 0.01, is not shown). The theoretical worst case, estimated with random neighborhoods, is approximately 0.5 for both measures. The results are for the yeast compendium data set. Publication 2 presents results also for mouse gene expression data. The figure is taken from Publication 2.

The continuity of the visualization, i.e., the ability to preserve original neighborhoods, was studied for the same set of methods. The results are shown in the right part of Figure 4.5. The SOM and multidimensional scaling (Sammon and non-metric MDS) were the best for preserving small ($k < 50$) original neighborhoods. Hierarchical clustering was by far the worst.

Trustworthiness of a visualization can be improved by discarding the least trustworthy data samples and analyzing them separately. This was done for the visualization of yeast genes shown in Figure 4.4. The trustworthiness of all of the analyzed methods improves when more and more samples are discarded from the visualization. However, the other methods do not reach even the starting point of the SOM before nearly one third of the data set has been discarded.

To conclude, the self-organizing map (SOM) was the most trustworthy except for the most similar gene expression profiles, where hierarchical clustering was the best. With regard to the continuity measure, the relative goodness of the methods depends on the data (see Publication 2 and also Venna and Kaski (2001)). In the comparisons with gene expression data, the SOM has performed well. Further proof for the feasibility of SOMs as the clustering and visualization measure for gene expression data is given by Gibbons and Roth (2002), who found the SOM to give more functionally coherent clusterings than hierarchical clustering algorithms.

4.6 Conclusions

In this chapter the self-organizing maps were used to study the function of yeast genes. At the same time, different distance measures suitable for gene expression data were compared; the correlation measure proved to be the best choice. The

learning metrics principle was used in conjunction with SOMs to further adjust the distance measures so that the visualized similarities would reflect the functional similarity even better. It was demonstrated that SOMs can find meaningful biological information from gene expression data sets. Furthermore, when the reliability of a SOM visualization in presenting similarities of genes was compared to that of other unsupervised methods, the SOM performed the best. To conclude, it is recommended that the SOMs in learning metrics are applied for gene expression data analysis.

Chapter 5

SOMs for grouping and visualizing retroviruses

In this chapter a variant of the SOM-approach introduced in the previous chapter is presented. Here the SOM-approach is extended to sequence data and used to study human endogenous retroviruses (HERVs). HERVs and transposable elements were introduced in Chapter 2. In this chapter their similarity relationships are studied, see also Publications 3-5.

The structure of this chapter follows that of the previous one. However, here more emphasis is put on discussions about the reliability of exploratory data analysis results. New methods are introduced for measuring the reliability of parts of a visualization display and for estimating the reliability of groups of samples manually extracted from a visualization display.

5.1 Problem setting

The taxonomy of HERVs is still incomplete. Currently the HERVs have been classified into approximately 40 HERV groups (Nelson et al., 2003; Mager and Medstrand, 2003; Gifford and Tristem, 2003), but the current classification is unable to categorize all new instances of HERVs detected in the human genome. Another problem is that in phylogenetic trees constructed from large HERV collections, some of the families are mixed with sequences from other families. This reflects the unfinished state of HERV classification. A new classification able to resolve these problems is needed. A better and clearer classification of the endogenous retroviral sequences will also help organize the "retrovirus universe", as most retroviruses are endogenous. In some recent publications the HERV groups are studied in more detail one by one (see Mayer and Meese, 2002; Jern et al., 2004; Yi et al., 2004). In contrast, in this chapter the SOM is used to study all HERVs together. This makes it possible to uncover new relationships between the existing HERV groups.

The aim of the SOM-based analysis of HERVs is to cluster them into groups having high sequence similarity within each group, i.e. by definition, originating from a common ancestor. Each such cluster will be one HERV group, or be formed of members from groups that were earlier thought to be separate. The clustering may also reveal groups of functionally similar HERVs, in those cases where the

HERVs have retained some function of the ancestral virus from which the HERVs are descending from. For example, the HERVs may have a functional gene that can code for a retroviral protein.

In this thesis the focus is on HERVs. Additionally, exogenous retroviruses and retrotransposons are analyzed together with HERVs in one application. The same approach can also be used to cluster protein sequences from various species, as demonstrated in (Kohonen and Somervuo, 2002). In such a setting groups of homologous proteins (protein families) should group together.

5.2 Distance measures for biological sequences

Distance measures that are used in this thesis to compare biological sequences are introduced here.

5.2.1 N-mer histogram presentation

A biological sequence can be converted into a vectorial representation by enumerating all its subsequences. Each component in the vector measures how often a specific N-mer, a contiguous subsequence of length N, is observed in the sequence. For DNA sequences, where the size of the alphabet is 4, a practical choice for N is 4, which results in 256-dimensional feature vectors. The feature vectors are normalized to unit length to make the presentation invariant to sequence lengths. Unfortunately, a practical N-mer length that does not produce prohibitively high dimensional feature vectors, is usually too short to capture all necessary sequential information. Still, the N-mer representation is useful because taking averages of samples is easy for vectors, whereas average of a group of sequences is not well-defined.

In this thesis the N-mer representation is used in conjunction with the Euclidean distance measure in training the sequence SOMs introduced later in this chapter. The vectorial representation is used in the organization phase where the averaging property makes the organization of the SOM easier. In the fine tuning phase of the SOM the HERVs are presented as sequences to utilize all information in the sequences.

5.2.2 Pairwise similarity scores

Pairwise sequence alignment algorithms, such as the Smith-Waterman algorithm (Smith and Waterman, 1981), can be used to measure the similarity of two sequences. The two sequences are aligned and then the mismatches and gaps in the alignment are counted. The final pairwise similarity score will be the sum over the aligned positions with matches giving a positive and gaps a negative increment to the score.

Any kind of a sequence alignment algorithm can be used to produce pairwise similarity scores. Popular alignment methods are, for example, the fast BLAST (Altschul et al., 1990) and FASTA (Pearson and Lipman, 1988) algorithms. They do not compute the full alignment, but resort to heuristics to speed up the computation. Of these two, FASTA is a bit slower and presumably more accurate for less similar sequences. It was used in Publications 3-5 to compare all data samples to each other. In Publications 6-7 the faster BLAST method was used to perform

database queries; the less accurate method was adequate because only a set of most similar database items was needed.

Tanimoto scaling

Pairwise similarity scores depend on the sequence length, because the score is a sum over all positions in the two sequences. When the sequence lengths of the analyzed sequences vary greatly, it may be a good idea to normalize the effect caused by the length. This can be done using Tanimoto scaling (Rogers and Tanimoto, 1960). The pairwise similarity score between sequences i and j is denoted by $s(i, j)$. These can be converted to Tanimoto similarities,

$$s_{Tan}(i, j) = \frac{s(i, j)}{s(i, i) + s(j, j) - s(i, j)}, \tag{5.1}$$

The Tanimoto similarities are between 0 and 1. The similarities can be further converted to the Tanimoto distance $d_{Tan}(i, j) = -\log s_{Tan}(i, j)$.

Other distance measures for sequence data

In addition to the measures used in this thesis, there are, of course, several other possibilities for measuring the similarity of biological sequences. Measures suitable for kernel methods, such as support vector machines (Cortes and Vapnik, 1995), have been proposed. These include the string kernel (Leslie et al., 2004) and Fisher kernel (Jaakkola et al., 1999), where the kernel function is derived from a hidden Markov model.

5.3 SOMs of retroviruses

Here the SOM is used for the task of grouping and visualizing a massive collection of HERVs. The visualization is used to refine the relationship among the HERV groups, and to detect potentially new groups (see section 5.1).

The traditional way to analyze HERVs is to use phylogenetic trees (PTs) that are based on a multiple alignment of the sequences. However, the high computational complexity of the multiple alignment step makes it difficult to use PTs for more than some hundreds of sequences. Heuristic methods exist to overcome this limitation, but the results may be biased. In contrast, the SOM is an algorithm capable of handling large amounts of data (Lagus et al., 2004). The SOM was also found to be the most reliable alternative among several visualization and clustering methods used for visualizing relationships of input samples (see Chapter 4). For these reasons, the SOM was selected as the method to analyze the large HERV collection.

5.3.1 Median SOM

The SOM can be used to order non-vectorial items such as DNA or protein sequences. The only requirement is that some distance measure is definable between the items. The SOM variant used to order non-vectorial data is called the median SOM (Kohonen and Somervuo, 2002). It resembles the batch-learning version of the plain SOM (Kohonen, 2001, 1996).

In the median SOM, the model of each map unit is defined as the *generalized median* of the input samples mapped into the neighborhood of the unit (Kohonen, 2001; Kohonen and Somervuo, 2002). The generalized median $\mathbf{m}$ is defined as the hypothetical data sample from which the sum of distances to the other elements $\mathbf{x}_j$ in a data set $\mathcal{D}$ is minimized, that is,

$$\mathbf{m} = \arg\min_{\xi} \sum_{\mathbf{x}_j \in \mathcal{D}} d(\mathbf{x}_j, \xi), \tag{5.2}$$

where $d(\mathbf{x}_j, \xi)$ is some distance measure defined between the $\mathbf{x}_j$ and ξ. In practice, the generalized median is often approximated by the *set median*; in the above equation ξ is then restricted to being an element of $\mathcal{D}$. The set median is an exact copy of one of the samples in the data set. If several samples in the data set satisfy Eq. (5.2), one of them is chosen randomly.

The best matching unit for each sample is selected practically the same way as in the basic version of the SOM (Eq. 3.5),

$$c = c(\mathbf{x}_j) = \arg\min_{i} d(\mathbf{x}_j, \mathbf{m}_i). \tag{5.3}$$

Here $d(\mathbf{x}_j, \mathbf{m}_i)$ is some distance measure defined between the sample $\mathbf{x}_j$ and the model $\mathbf{m}_i$. In the adaptation step, the new value for the model $\mathbf{m}_i$ is determined as the set median of those input sequences that were mapped to the said unit, or to its neighborhood on the SOM grid.

The SOM has been previously applied to problems in biological sequence analysis. In several cases the sequences are first transformed into vector representations and a normal vectorial SOM is used, for example, by Kanaya et al. (2001); Mahony et al. (2004) and Wang et al. (2001) to study codon usage, by Abe et al. (2003, 2006) to study 3- or 4-mer frequency patterns in different species or genomic areas and by Ferrán and Ferrara (1991); Hanke and Reich (1996) and Yang and Chou (2003) to cluster protein sequences. In contrast, in this thesis SOMs that can directly use the biological sequences are applied. The median SOM algorithm has previously been applied to clustering of similar protein sequences by Kohonen and Somervuo (2002). This thesis presents the first real biological application of the method and demonstrates that new biological knowledge can be gained through the median SOM analysis. SOMs that can directly use biological sequences have later been used by Mahony et al. (2005a,b, 2006) to find transcription factor binding motifs. Their SOMBRERO algorithm differs from the median SOM in that the model of each SOM unit is a probabilistic presentation of the most likely nucleotides of a transcription factor binding site, and not a simple sequence like in the median SOM.

5.3.2 SOM of transposable elements and retroviruses

In Publication 3 the median SOM approach is applied to grouping and visualizing a small collection of retrotransposon consensus sequences (HERVs and LINEs[1]) together with genome sequences of exogenous retroviruses. The aim of the study was to show that the median SOM is able to separate different types of sequences from each other (LINEs separate from the viruses and so on) and to group similar sequences together (like Class II HERVs together with betaretroviral retroviruses).

[1]Long interspersed repeat sequences (LINEs) are one type of retrotransposons

The results verified that the median SOM is able to do this. The SOM visualization and more information about the biological results can be found from Publication 3. The results show that a completely data-driven grouping is able to reflect same kinds of relationships as more traditional phylogenetic taxonomies.

5.3.3 SOM of human endogenous retroviruses

After the proof of concept application (grouping of clean consensus sequences in Publication 3) the median SOM was applied to "real data" in Publications 4-5; it was used to analyze all HERVs that can be automatically detected from the human genome. The aim of the study was to characterize the HERVs and refine the existing classification. The study was the first to analyze all HERV sequences simultaneously. Previous approaches using phylogenetic trees have used only some subsets of the HERV sequences, due to the high computational complexity of the PT algorithms. The *HERV SOM* visualization is shown in Figure 5.1.

Biological results

The SOM finds the division of HERVs into the standard HERV classes I-III: each class is localized to its own area on the display. Furthermore, the previously characterized HERV groups can also be detected with the SOM. The class distribution of each group is focused on a set of nearby map units. Only few HERV groups spread out more, or mix with other established groups, reflecting the uncertainty in the current HERV classification.

In a comparison to phylogenetic trees (PT) constructed from representative subsets of the HERV sequence collection (500 sequences) the groupings from SOM and PT were found to be similar. However, the SOM detected biologically interesting sequence groups that were not visible in the phylogenetic trees. One is a group where three HERV groups, previously thought to be separate, mix together (marked with number 1 in Figure 5.1). Furthermore, SOM detects several groups of unclassified sequences (marked with '?' in the figure), one of which turns out to be a group of chimeric HERV elements and another to be a group of epsilonretroviral sequences. Epsilonretroviruses have not been previously detected in humans.

The area where sequences from three HERV groups, ERV9, HERVW, and HUERSP3, are mixed together is called "Area 1" and is analyzed more closely in Publication 5. A PT is constructed from all HERVs classified to these three HERV groups. The tree confirms the observation from the SOM: the sequences in Area 1 truly form a new separate group.

5.4 Reliability of visualization results

The SOM-based analysis is complemented with estimation of the reliability of the results. In addition to measuring the reliability of the whole visualization, measures are proposed for estimating the reliability of different areas on the visualization display and of groups of samples extracted manually from the visualization for further analysis.

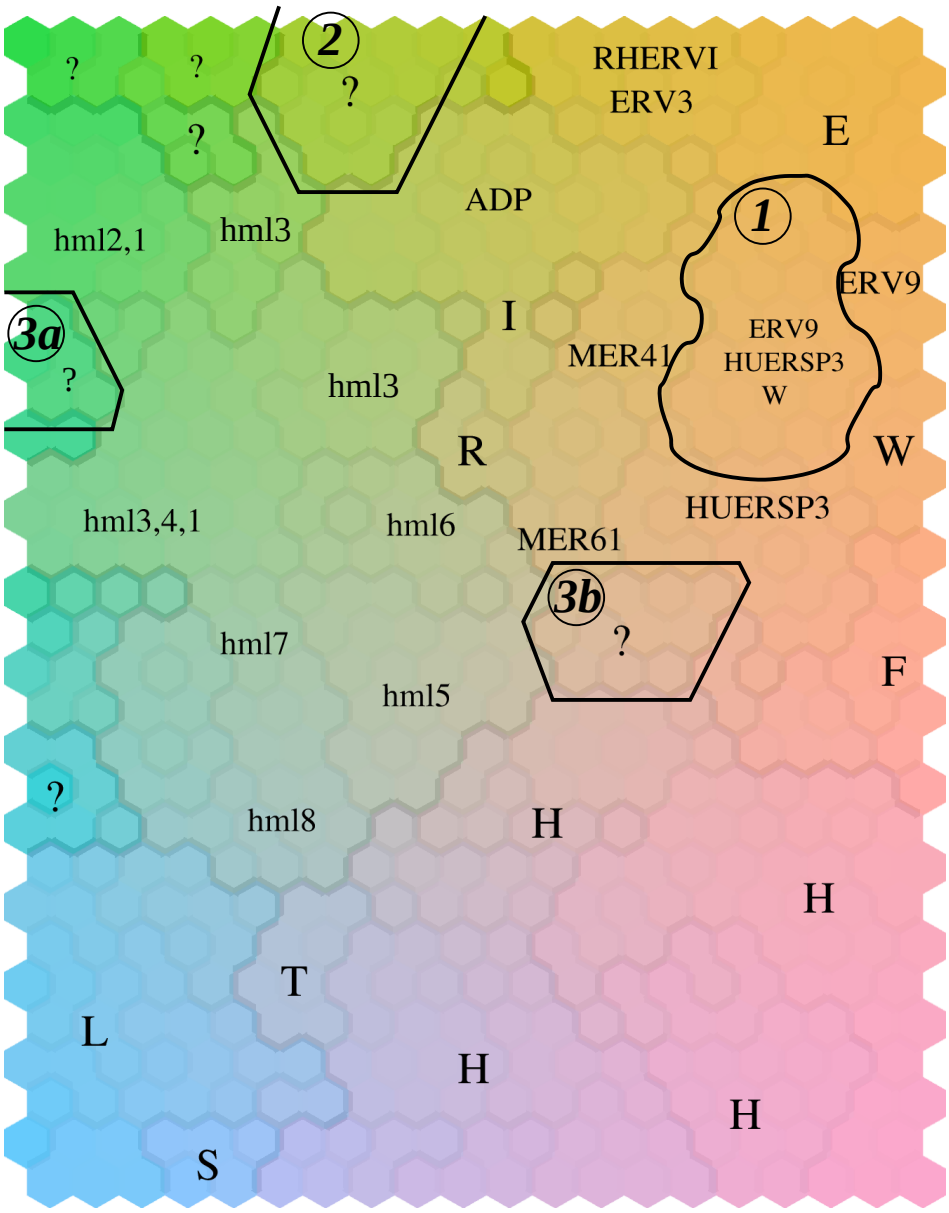


Figure 5.1: A U-matrix visualization of the *HERV SOM*. The distance between the neighboring units is denoted by the grey shade of the edge between them and also by the change in the shade of color between the neighboring nodes (Kaski et al., 1999). The labels, referring to current HERV classes, are manually assigned descriptions for the map areas. Question marks denote areas that contain unclassified HERVs. The continuous dark borders separate dissimilar areas from each other. For example, in the middle of the map the sequences in the HERVH group (denoted by the label H) are separate from the sequences in the HML5 group (denoted by the label hml5). The three circled areas are interesting clusters that have been analyzed more closely in Publication 5: in area 1 HERVs from three HERV groups are mixed together, area 2 contains chimeric HERVs, and areas 3a and 3b contain epsilonretroviral HERVs.

5.4.1 Trustworthiness of the median SOM

Similarly as in the case of SOMs for gene expression data, the trustworthiness of the SOM display was compared to that of the competing method, in this case the phylogenetic tree. Due to the high computational complexity of the multiple sequence alignment step, the PTs used in the comparison were computed from a pairwise similarity matrix and not from a multiple alignment; this effectively reduces the method to hierarchical clustering.

The results obtained in Publication 5 show that at its best the SOM outperforms the PTs, but that the PTs perform better than the SOMs do on average. However, when the number of sequences in the neighborhood, k, increases to a number used when analyzing the SOM in practice (40 and above), the difference between the SOM and PT curves is small. Hence, the SOM is a reasonable although not clearly the most trustworthy choice for large scale sequence analysis.

5.4.2 Reliability of areas of the SOM visualization

In addition to measuring the reliability of a whole visualization, it makes sense to evaluate the reliability of different parts of the visualization separately. If some areas of the display are found to be less reliable, then the analysis of the visualization can be focused on the more reliable areas and new visualizations computed for the samples in the unreliable areas. Here two measures are presented for measuring and visualizing the relative reliability of each location on the SOM display.

Reliability by bootstrapping

The reliability of the SOM clustering and visualization is estimated by evaluating how similar the SOM display is in bootstrap repetitions of the SOM. The SOM result is called stable if the displays are always similar. In practice, the stability is measured by counting how frequently a pair of sequences appear nearby on the SOM display in the bootstrap repetitions. Stability of groupings presented by individual map units in the visualization are derived as averages over the sequences in that map unit.

Stability of the SOM is estimated separately for each pair of sequences. Only the immediate neighborhood on the map (the same map unit and its bordering units) is considered, but other choices of neighborhood size are possible as well. The frequency $f_{i,j}$ of samples i and j appearing as neighbors on the bootstrap maps is counted and a measure for the reliability of each map unit is computed as the average stability among the pairs of sequences in that unit:

$$M_B(u) = \frac{1}{N_u(N_u - 1)} \sum_{i \neq j, i, j \in \text{ unit } u} f_{i,j}, \tag{5.4}$$

where N_u is the number of sequences in the unit u. A measure similar to Eq. 5.4 can be computed for larger groups of sequences as well, for instance for clusters of SOM units.

A visualization of the reliability scores (Eq. 5.4) for each map unit in the HERV SOM (the one in Fig. 5.1) is presented in Figure 5.2 (left). The visualization reveals reliable clusters and areas where the visualized similarities are unreliable. The overall reliability of the visualization is reasonably good. The average reliability score of the map units is 0.52, which is much better than a average score for random

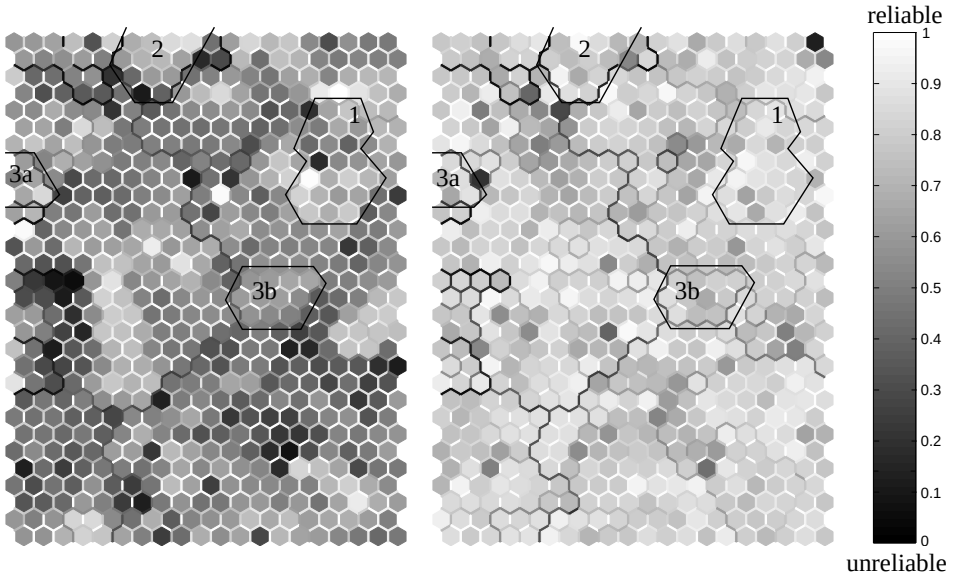


Figure 5.2: Reliability of the map shown in Figure 5.1. Left: reliability by bootstrap. Right: reliability by the trustworthiness measure. The gray scale is the same in both figures. However, the absolute values of the reliability scores from the two measures are not directly comparable. The gray shade differences are meaningful only within one display. The figures are taken from Publications 4 and 5.

assignment of the sequences into the SOM units (about 0.22). The biologically interesting areas (circled in the display in Fig. 5.2 (left)) are not the most reliable ones but still relatively reliable (around 0.65 on average) and thus worth analyzing further.

Bootstrapping has previously been used in conjunction with clustering to assess the the stability of the clustering results with respect to sampling variability. For example, in (Bhattacharjee et al., 2001) the consensus matrix (see section 3.5.2) from bootstrap repetitions of a partition-based clustering was visualized to provide support to a hierarchical clustering result. Monti et al. (2003) used the consensus matrix as a input data to a consensus clustering. They also propose a statistic to summarize the stability of a cluster (see section 3.5.2). Their cluster consensus measure is very similar to the measure in Eq. 5.4 (and identical to the R-index proposed by McShane et al. (2002)); the difference is in the definition of the frequency $f_{i,j}$. They use the frequency of two samples being in the same cluster, whereas above, $f_{i,j}$ is defined as the frequency of two samples appearing *nearby on the SOM display*, i.e., the visualization aspect of the SOM and the fact that clusters are usually defined as sets of SOM units are taken into account. Furthermore, the reliability scores are used to *visualize* the reliability in different areas of the SOM display. This is a novelty over previous approaches.

The bootstrapping procedure has previously been applied to self-organizing maps by de Bodt and Cottrell (2000). The article describes significance tests for the quantization error and for the stability of neighborhoods on the SOM. Here the aim is not at significance tests but on the ability to look at the stability of the neighborhoods *in each map unit separately*.

It should be noted that also other resampling schemes besides bootstrapping

can be used to estimate the stability of a clustering with respect to sampling variability. For example Levine and Domany (2001) and Ben-Hur et al. (2002) use subsampling to produce a set of perturbed data sets. Each of the subsets is clustered and the obtained clustering is compared to the one obtained from the complete data set and a stability score is derived for the complete clustering. In these works the aim is to find the set of parameters (for example the number of clusters) that produces most stable clusterings. The aim is different from the one taken in this thesis, where the resampling is used to evaluate each SOM unit separately and, most importantly, to visualize the relative reliabilities of the units.

Reliability by trustworthiness measure

The trustworthiness measure, introduced in section 3.5.1, is extended here so that the trustworthiness of different areas on the display can be measured. An area on a display is considered trustworthy if all samples close to each other in this area of the display can be trusted to have been proximate in the original space as well.

When measuring the trustworthiness on the SOM display it must be decided how to define the neighborhood of each sequence s. A practical choice would be to select the sequences from the same map unit and its neighboring units up to a pre-selected radius as done in the bootstrap measure above. The problem with this approach is that it disregards clusteredness in the data as neighboring map units may belong to different clusters (distance between neighboring units is large). Furthermore, the number of neighbors would vary and could lead to low quality trustworthiness estimates when the number is small. To get reliable estimates and to take into account the clusterings visible on SOM displays, the sequences are selected from close-by map units in the order of their distance from the unit where s is. The distance is computed along the minimal path on the map grid where the distance of neighboring units is defined as their distance in the data space (the U-matrix distance which measures visual closeness). Sequences are collected until their number equals or exceeds a preselected number k. Similarly as in the case of the trustworthiness measure, this number k should be close to the number of sequences a person looking at the SOM is likely to consider similar. Here it is assumed that sequences that are located in the same or in neighboring SOM units are considered similar.

The untrustworthiness of the map display is estimated at each map unit u separately. Denote the set of sequences within unit u by I_u. A measure of the untrustworthiness of a map unit is computed as an average over the N_u sequences in the unit by

$$M_T(k,u) = \frac{1}{N_u} \sum_{s_i \in I_u} \sum_{s_j \in U_k(s_i)} (r(s_i, s_j) - k) \;, \tag{5.5}$$

where the untrustworthiness of one sequence s_i is computed similarly as in the trustworthiness measure (Eq. 3.11). The untrustworthiness scores are converted to trustworthiness scores with the transformation $1 - M_T(k,u)/A(k)$, where $A(k)$ is used to scale the values between zero and one.

The trustworthiness values of the HERV SOM (the one in Fig. 5.1) are visualized on the SOM display in Figure 5.2 (right). In a white (or light gray) unit the average trustworthiness of the sequences in the unit is very good, i.e. there are few sequences in the SOM neighborhood that are not in the neighborhood of the sequences in the original data space. As can be seen from the image, the whole

SOM is reasonably reliable. On average the trustworthiness value of a SOM unit is around 0.8.

Comparison of the two approaches

Both of the measures above serve the same purpose. The idea is to reveal which parts of the visualization display are more reliable than others. The first, bootstrap-based measure, focuses on the clustering ability of the SOM. If the set of samples in a SOM unit is always grouped together in the bootstrap repetitions, then the set of sequences in the unit is considered a reliable group. The bootstrap-based method does not quantify the observed vs. original similarities of the samples, but looks only at the repeatability of clusterings on the SOM. The second trustworthiness-based measure focuses on the visualization ability of the SOM. Those areas of the display where the original similarity relationships are conserved the best are the most reliable.

The trustworthiness-based measure can be considered to be an improved version of the earlier measure. It takes the distances of the units along the SOM grid into account and always computes the measure for a fixed-sized neighborhood on the SOM display. These same improvements could, in principle, be introduced also to the bootstrap-based measure by defining the neighborhood of sequences on the SOM differently in the computation of the co-occurrence frequency $f_{i,j}$. Then the difference between the two measures remains in the distinction whether clustering or visualization ability is more interesting.

5.4.3 Reliability of groups of samples extracted from a visualization

The SOM visualization can be used to extract interesting groups of mutually similar sequences for further analysis. The U-matrix visualization shows with gray shades how close the models of neighboring map units are, and clusters can be defined as sets of close-by units. The exact borders of the cluster areas on the visualization are selected partly subjectively. The decisions are based on various labelings describing the contents of each SOM area, on the U-matrix visualization, on the reliability visualization, and on all background knowledge the analyst has about the data set. Finally, a set of SOM units is selected and the sequences with them are extracted for further analysis. This group of extracted sequences is denoted with $\mathcal{C}$. Before the sequences in group $\mathcal{C}$ are analyzed further it should be checked that they really form a reliable cluster.

In this thesis two new measures are proposed for evaluating the reliability of groups of sequences ($\mathcal{C}$) extracted from one specific SOM visualization. The measures are based on the bootstrap method (see section 3.5.2) and the assumption is that if a group of sequences always appear together in SOMs constructed from sampled data sets, then those sequences represent a true group (cluster) in the data. Measures of cluster stability (Levine and Domany, 2001; Ben-Hur et al., 2002; Monti et al., 2003) cannot be directly applied to the SOM because they apply to a complete clustering, whereas here the primary interest is in single manually extracted clusters that are formed of sets of close-by map units on the SOM display. The measures also differ from the bootstrap-based stability measures used in clustering in that they take the orderedness of the SOM display into account.

Two measures, the compactness and purity of the cluster, are proposed for the task of estimating deviations from a perfect clustering (on the SOM display) in bootstrap repetitions of the SOM. The data set visualized on the SOM that was used to *define* the group $\mathcal{C}$ is resampled and a new SOM is computed for each resampled data set. Compactness measures how close together the group of sequences is on the bootstrap SOM. Purity, on the other hand, measures how many foreign sequences are mapped to the same area as the interesting group. The purity of the cluster is analogous to precision in information recall, and false alarms in detection theory. The proposed measures are new, and have not been previously used in this form. The *concept* of compactness is, however, not new. It is commonly used to evaluate the quality of a clustering result (without resampling). The measure is then the compactness or homogenity of the cluster in the original data space. For a review of cluster quality measures see Handl et al. (2005). In contrast, here the compactness is defined on the SOM visualization display and not in the original data space.

The compactness and purity (C_b and P_b respectively) of the selected group of sequences are measured in each bootstrap repetition b and a sampling distribution is obtained for each measure.

Compactness and purity are measured as functions of the varied number of samples in the group $\mathcal{C}$, to take into account possible substructure in the group. First the measures are evaluated for the whole group, then the sequence which most worsens the measure is removed from the group and the measure is evaluated again. $\mathcal{C}_k$ is used to denote the group of sequences still remaining in the set after k sequences have been removed. The removal of the worst sequence and re-evaluation of the measure is repeated until no sequences are left. The removal of sequences is carried out separately for the two measures. For the purity measure this removal process is optimal. Usually also the compactness improves steadily with this removal procedure which is a greedy approximation of an optimal procedure.

More formally, the compactness $C_b(k)$ after k removals is defined as

$$C_b(k) = 1 - \frac{\max_{i,j \in \mathcal{C}_k} d(u(i), u(j))}{D_{max}}, \tag{5.6}$$

where $u(i)$ is the location of the map unit containing the sequence i, d denotes the Euclidean distance of the units along the SOM grid (distance between the centers of bordering units is one), and D_{max} is the maximum distance between units on the SOM. So, the compactness measures how close together the sequences in group $\mathcal{C}$ are on the bootstrap SOM. If the distance $d(u(i), u(j))$ is very large for some pair of sequences, it is clear that the group $\mathcal{C}$ is not located in a set of nearby units.

The purity $P_b(k)$ after k removals is defined to be

$$P_b(k) = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \frac{\sum_{j \in \mathcal{C}_k} InSameUnit(i, j, b)}{\sum_{j=1}^{N} InSameUnit(i, j, b)}, \tag{5.7}$$

where InSameUnit(i, j, b) is the indicator function that returns 1 if i and j are in same unit on the SOM of the bootstrap sample b, and otherwise zero. The purity is measured as a sum over the sequences in group $\mathcal{C}_k$. For each sequence i it is computed how large a fraction of the sequences located in the same unit as i are also from the group $\mathcal{C}_k$. Measures somewhat similar to the purity measure have also

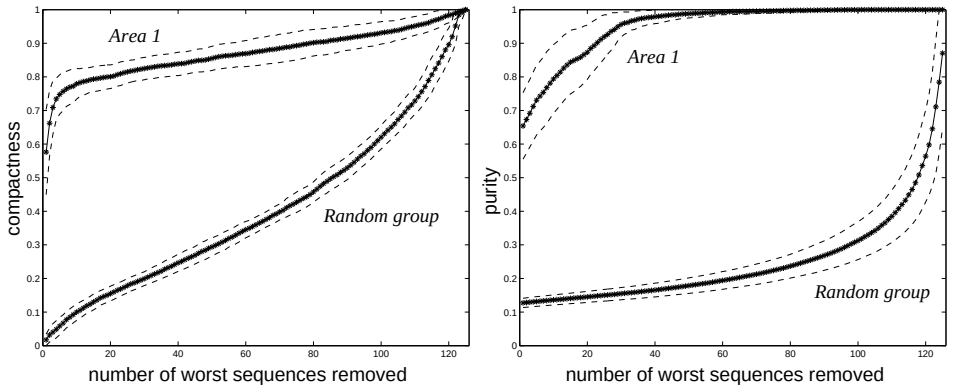


Figure 5.3: Compactness (left) and purity (right) for the group of sequences in Area 1. The solid curves represent the average compactness (purity) over the 100 bootstrap maps, the dashed lines the mean ± standard deviation of the distribution. Note that the compactness value of one is not attainable in practice for large groups of sequences as they usually do not fit into a single map unit. For reference, the compactness of the group of sequences in Area 1 in the SOM computed from complete data (that in Fig. 5.1) is 0.86. The figures are taken from Publication 5.

been proposed earlier. Bryan (2004) and Ben-Hur et al. (2002) present a measure for a complete clustering: the fraction of sample pairs that are clustered similarly in a bootstrap clustering as in a reference clustering. In the purity measure, instead of a global evaluation over all clusters, the situation is looked from the point of view of a single cluster (the set of SOM units that contain members of group $\mathcal{C}_k$): how large a fraction of sample pairs in that cluster were also together in the reference clustering (i.e. both partners of the pair are in the set $\mathcal{C}_k$).

The compactness and purity measures are presented as curves that are a function of the number of sequences removed. The curves are analogous to receiver operating characteristic (ROC) curves. A sharply rising curve is better that a nearly linear one. A sharp incline tells that the group of sequences is very homogeneous. Removing merely the few worst sequences brings the group's performance to the highest level.

The "Area 1" in the HERV SOM is interesting because sequences from three HERV groups are mixed together in that area (see Section 5.3.3). The reliability of the group of sequences extracted from Area 1 was studied to verify that the observed mixed group is a true cluster in the data. Here the cluster is analyzed as an example on how compactness and purity can be used in the analysis of a group of sequences extracted from the SOM display. The compactness and purity of this cluster are very good compared to an equal-sized randomly sampled control set (see Fig. 5.3). Both compactness and purity rise rather quickly to the reasonable level of 0.86 and 1, respectively; these figures are measured from the original SOM (shown in Fig. 5.1) used for *defining* the group, and hence represent a kind of best possible reasonable values. The results suggest that the cluster is not an artifact, but really exists in the HERV sequence collection. The biological analysis of the cluster (see Section 5.3.3 and Publication 5) agrees with this statement.

5.5 Conclusions

In this chapter the median SOM, a self-organizing map for sequential data, was used to group and visualize a comprehensive collection of human endogenous retroviruses and, at the same time, to study the HERV classification. As a result of the SOM analysis the classification of three HERV groups was redefined. The SOM also detected a new, previously undiscovered group of epsilonretroviral HERVs. These results show that the SOM is able to find new information about HERVs that have previously been studied with phylogenetic trees.

The SOM analysis was complemented with estimates on the reliability of different parts of the SOM visualization. Two new measures were presented for estimating the reliability of each map unit on a SOM display. A reliability visualization is obtained when the reliabilities of the units are shown on the SOM display. The reliability visualization can then be used to focus the analysis on the most reliable areas of the SOM display. Of the two new measures the trustworthiness-based one is preferred. It can be applied also to various other visualization methods. For example, all individual points in a projection or all leaves and inner vertices in a hierarchical clustering tree could be colored according to their trustworthiness.

The SOM visualization is often used to select interesting groups of mutually similar sequences for further analysis. The selection is usually done manually and is based on several visualizations of the SOM (U-matrix, reliability, textual labels) and all sorts of prior information. It is advisable to check the reliability of the groups of sequences/samples before further analysis. Here two new measures for estimating the reliability of such groups were presented.

Chapter 6

Hidden Markov mixture model for estimating HERV activation

This chapter introduces a new hidden Markov model (HMM) based probabilistic model that is designed to solve a specific biological question: Which individual HERV sequences are active? The chapter begins by defining the biological problem and then the new model and a heuristic alternative to it are described. The rest of the chapter is devoted to the experimental results from Publications 6 and 7.

6.1 Problem setting

The thousands of HERVs in the human genome are a huge potential source of active sequences, yet the HERVs are much less studied than genes. Generally, from the fundamental research point of view, it would be interesting to understand whether these repetitive elements are activated and, in particular, when and where they are activated. The main interest in studying HERV activation is, however, the connection to diseases; some HERVs are expressed in diseases, but it remains unclear whether the observed expression is causing the disease (see Section 2.4.3 for a longer discussion). Thus, it is vital to understand the details, causes and effects of HERV activation. The first step is to study the potential for activity of all the HERV integrations in the human genome (Publication 6). In the second stage the expression profiles of HERVs (over a set of tissues) are studied to get a deeper understanding of the function of HERVs; this work was started in Publication 7.

Methods used so far for studying HERV activation were already reviewed in chapter 2. Briefly, HERV activity has been observed but due to limitations of the laboratory methods, it is not known which individual HERVs are active. Current HERV activity measurement techniques estimate the activity of a whole HERV group together, see, e.g., Seifarth et al. (2005); Hu et al. (2006); the only exceptions so far are Stauffer et al. (2004) where a small test for individual HERVs of one group was done with a heuristic method and Kim et al. (2005) where HERVs were sought from gene mRNAs but activities were not compared across HERVs. However, the genomic locations of active HERVs are needed in order to uncover

Figure 6.1: The EST matching problem. The sequences on the left are the HERVs. Their sequences are nearly identical; the darker gray boxes denote locations where the HERVs differ. The topmost HERV is transcribed and subsequently converted to an EST sequence. Note that the EST is truncated and contains one sequencing error. The EST is matched against the HERVs in order to determine where it came from. This is difficult, because the EST matches three HERVs equally well (the 1st, 4th and 5th HERV from the top).

the control elements causing the activity in each case (see Section 2.4.3). So, a method able to measure the activity of each individual HERV locus in the genome is needed.

In this thesis an *in silico*[1] approach for estimating the activities of HERVs is presented. Evidence of HERV activation can be found from the large database of expressed sequence tags (ESTs), the dbEST (see Section 2.5). The ESTs, introduced in Section 2.3.3, provide and abundant, albeit noisy, information source about expressed genomic elements. The EST database contains mainly gene transcripts but also transcripts of other active sequences, such as HERVs. Special features of the ESTs are the following (Section 2.3.3): they are truncated versions of either the beginning or end of the mRNA sequence of a gene[2], they are not exact replicates of the mRNA, but contain sequencing errors (mismatches, deletions, insertions), and the ends of the ESTs are of lower quality (more and more sequencing errors).

In principle, the ESTs can be used to measure HERV activity by counting how many ESTs match each HERV. Active HERVs will have numerous EST matches and inactive HERVs none. However, unambiguous matching of ESTs to HERVs is very difficult. This is because the noise level (sequencing errors) in ESTs can be larger than the sequence differences between two HERVs. Then it is difficult to say which of the two nearly identical HERVs was the source of the EST. This is the *EST matching problem* (see Fig. 6.1).

In this thesis a solution to the EST matching problem is presented in the form of a biologically motivated probabilistic model that learns the relative activities of the HERVs from EST sequence data.

6.1.1 HERV and EST sequence data

This subsection introduces the data sets used in Publications 6 and 7. In Publication 6 the activation of *all* HERV sequences was studied. A collection of over 3000 HERV sequences was obtained automatically from the human genome using the RetroTector program (see Section 2.4.4). In addition to the full HERV collection, two well-chosen smaller subsets of HERVs were studied in the publications. The first set contains 181 sequences from three HERV groups: HERVW, HML2,

[1] *In silico* means "by computers" and is the opposite of the two terms: *in vitro* (in the test tube) and *in vivo* (within a living cell), that refer to laboratory techniques.

[2] Sometimes the EST may even come from the middle of the mRNA. However, most of the ESTs are sequenced from the beginning or end of the transcript.

and HERVE. The groups were selected based on previous studies where they were reported to be active, see, e.g., Seifarth et al. (2005); Forsman et al. (2005). The second, smaller set consists of sixty HML2 sequences, whose activity profiles were estimated and analyzed in Publication 7. The HML2 group was selected because it has the most potential for containing active HERVs and because it was found to be active in Publication 6.

ESTs matching the HERVs were retrieved from the dbEST using BLAST (Altschul et al., 1990). BLAST (basic local alignment search tool) is a fast algorithm for computing pairwise sequence similarity and is commonly used in database queries. When constructing the collection of retroviruslike ESTs, each HERV sequence in its turn was used to query the dbEST. Details on EST to HERV matching and EST data preprocessing can be found in Publications 6 and 7.

In Publication 6 the HERV data was pruned by removing those HERVs where a non-viral[3] part of the HERV had EST matches. The removal left out HERVs where activity comes from suspected non-retroviral areas. The pruning made it possible to focus on expression of retroviral origin. In the end the HERV collection contained 2450 HERVs.

6.2 Generative model for HERV expression

Publication 6 introduces a probabilistic model for handling the uncertainty in the EST to HERV matching.

The model has been designed based on the following assumptions about how ESTs are generated from HERVs: (i) EST transcription starts at some point of the HERV sequence; (ii) the EST sequence follows the HERV sequence, but (due to sequencing errors) can contain mismatches between the EST and HERV nucleotide, and can skip HERV nucleotides or insert new ones; (iii) lastly, the end of the EST sequence is of lower quality and does not resemble the HERV sequence.

The new model is a generative mixture model for the set of EST sequences. The mixture components are the HERV sequences, i.e., the possible sources of ESTs. The model mimics the actual generation of the ESTs from the set of HERVs.

6.2.1 Hidden Markov mixture model

Each mixture component in the generative model is a hidden Markov model (HMM) for ESTs from a particular HERV (see Fig. 6.2). Each component HMM is similar to the profile HMM (Section 3.4 and Krogh et al. (1994)), with the exception that it is possible to jump from the start state to any of the match states and from any match state either to the end or to a special EEMIT state that is used to emit the low quality end of an EST. The match states, one for each position of the HERV sequence, can either emit the nucleotide in that position of the HERV sequence (with probability p_t) or one of the other nucleotides (with probabilities $(1 - p_t)/3$). To summarize the model structure, each component HMM generates data that roughly matches a subsequence of the source HERV, but with mismatches, insertions, deletions, and a low-quality end part.

[3]Non-viral meaning here DNA that the RetroTector has not annotated as part of a virus gene or LTR. The HERV chain from RetroTector may contain these non-viral stretches in between viral parts. They may be either heavily mutated virus sequence parts or non-viral DNA that has ended up in the middle of a virus due to genomic rearrangements.

The complexity of the model is constrained by sharing parameters. The parameter p_t is shared between all match states in the model. Similarly, all the EEMIT and insert states share parameters: they emit nucleotides using the same distribution. The emission parameters of the model are summarized in Table 6.1. The transition parameters are also shared throughout the mixture (see Fig. 6.2): all the basic blocks of all the sub-HMMs are identical except for the preferred nucleotide in the match state that depends on the HERV sequence. The transition parameters of the model are summarized in Figure 6.3.

The mixture model can be interpreted as one large HMM where the first transition chooses one of the N HERV-specific sub-HMMs (see Fig. 6.2). The transition parameters of the first transition are the mixture weights $\mathbf{a} = \{a_1, ..., a_{N_{herv}}\}$ and correspond to the activity estimates for the HERVs. The probability distribution for one data item (the i:th EST sequence) given the model parameters $\mathbf{a}$ (mixture weights) and θ (transition and emission parameters) is

$$p(\mathrm{EST}_i|\mathbf{a},\theta) = \sum_{j=1}^{N_{herv}} p(\mathrm{EST}_i|j,\mathbf{a},\theta)p(j|\mathbf{a},\theta) = \sum_{j=1}^{N_{herv}} a_j p(\mathrm{EST}_i|\mathrm{HERV}_j,\theta), \quad (6.1)$$

Due to parameters sharing the parameters θ are the same for all the N_{herv} sub-HMMs, the only difference between them is the HERV sequence, denoted by HERV_j in the equation.

The basic HMM training procedure, the Baum-Welch algorithm, can be used to learn the whole mixture. Batch update rules for the shared transition and emission parameters and mixture weights can be derived similarly as in the case of a full HMM model. The parameter sharing, however, reduces the number of parameters drastically and makes it possible to learn the parameters of the complex model with the limited number of data available. The computational complexity of the model is discussed in Publication 6.

6.2.2 Estimating HERV expression profiles

The HMM mixture introduced above can be extended for the task of estimating HERV expression profiles. First, a separate model is learned for each condition, using a set of ESTs specific to that condition. Then the relative activity distributions of the HERVs from different conditions are combined to form a HERV expression data matrix (see Fig. 6.4). However, it is not immediately clear how activity estimates from different conditions should be scaled before they are combined. There are two ways to do the scaling: 1) No scaling is used. In this setting it is assumed that each EST set, irrespective of its size, is a *representative sample* of all HERV-derived mRNAs in the condition. If this is true then the relative activity distributions from different conditions are directly comparable. 2) The relative activity distribution of each condition is scaled by the number of ESTs available from that condition. This transformation makes the activity estimate of a HERV more directly proportional to the actual number ESTs coming from that HERV; the activity value of the HERV can be seen as a *probabilistic EST count*. In this setting it is assumed that the size of the EST set is relevant. In Publication 7 the second approach was used.

Another problem in the estimation of HERV activity profiles is the cross-talk arising from HERVs not included into the studied set. Let us consider a case where the HML2 HERVs are studied. These HERVs are used to retrieve a set of

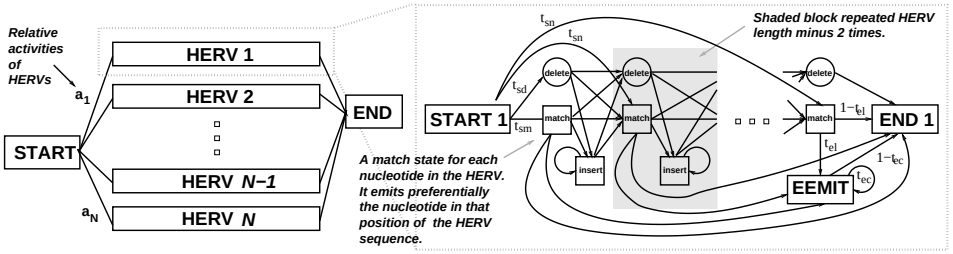
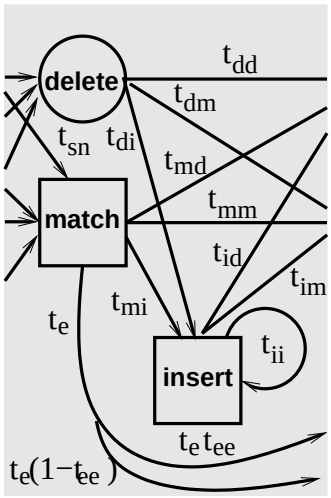


Figure 6.2: The structure of the HMM mixture. The model is constrained by sharing parameters. The shaded box is the basic block of the sub-HMM; close-up of it is shown in Figure 6.3. The block is identical in all sub-HMMs, only the emission distributions of the match state vary.



t_{sn}	start state to match states 2...L
t_{sd}	start state to delete state 1
t_{sm}	$1-(L-1)*t_{sn}-t_{sd}$
t_{dd}	delete state to delete state
t_{di}	delete state to insert state
t_{dm}	$1-t_{dc}-t_{di}$
t_{id}	insert state to delete state
t_{ic}	stay in the insert state
t_{im}	$1-t_{id}-t_{ic}$
t_{md}	match state to delete state
t_{mi}	match state to insert state
t_e	match state to end/EEMIT
t_{ee}	to EEMIT state and not to END
t_{mm}	$1-t_{md}-t_{mi}-t_e$
t_{el}	last match state to EEMIT state
t_{ec}	stay in the EEMIT state

Figure 6.3: Transition parameters of the HMM model. Left: A close up of the basic block of the sub-HMM. The block is repeated HERV length minus two times in each sub-HMM, but actually also the first transitions (like between the first and second match state) are the same as shown here for the basic block. The emission distribution of the match state varies between blocks (and sub-HMMs), according to the HERV sequence each sub-HMM corresponds to, see Table 6.1. Right: descriptions of the transition parameters. Note that the transitions to match states can be computed from the other parameters. Thus, there are effectively only 12 transition parameters for the whole model.

	A	C	G	T
all insert states	e_A	e_C	e_G	e_T
all EEMIT states	e_A	e_C	e_G	e_T
match state when s_i = A	p_t	$\frac{1}{3}(1-p_t)$	$\frac{1}{3}(1-p_t)$	$\frac{1}{3}(1-p_t)$
⋮			⋮	
match state when s_i = T	$\frac{1}{3}(1-p_t)$	$\frac{1}{3}(1-p_t)$	$\frac{1}{3}(1-p_t)$	p_t

Table 6.1: Emission parameters of the HMM model. All the insert states and all the EEMIT states have the same emission distribution. For the match states the emission distribution depends on the HERV sequence. If the i:th nucleotide of the HERV sequence s is A then the emission distribution for the i:th match state is as shown in the table for the case s_i = A.

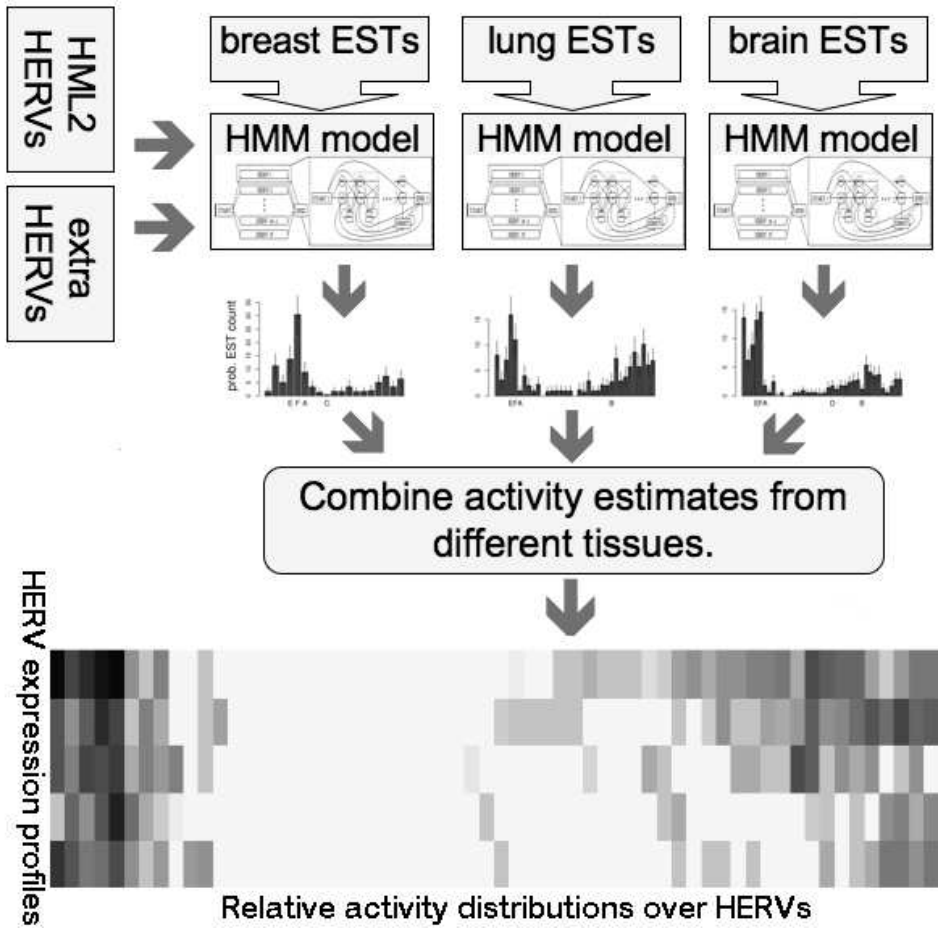


Figure 6.4: The process of estimating HERV expression profiles.

ESTs from the database. However, the set may contain ESTs that are actually originating from a non-HML2 HERV. This means that there is cross-talk between the HERV groups; ESTs coming from a HERV in one group also match HERVs from another group. To reduce the errors caused by the cross-talk some elements from other HERV groups (extra HERVs) should be included to the HERV set. This is done to ensure reliable activity estimates for the interesting HERVs (HML2 HERVs). The extra HERVs will capture the ESTs that are not HML2-derived.

The set of extra HERVs should be as representative as possible. In an ideal situation it would contain all the HERVs that do cause cross-talk. In practice, HERVs that may cause cross-talk can be detected by comparing the ESTs retrieved using HML2 HERVs to all HERVs. HERVs that match the ESTs very well potentially cause cross-talk. Thus, the set of extra HERVs should be selected among these. If it is not computationally feasible to include all of them, then a representative subset may be obtained by selecting some HERVs from all HERV groups. This works because HERVs within a family are so similar that they match each other's ESTs very well.

6.2.3 A heuristic alternative

A straightforward alternative to the HMM mixture is to neglect any cross-talk between the HERVs. Their activities can then be estimated simply by the number of BLAST hits. The BLAST activity of a HERV is the number of ESTs matching that HERV better than any other HERV. A similar BLAST approach was used by Stauffer et al. (2004) for a tiny data set containing only intact HERV sequences.

6.2.4 Estimating the reliability of the activities

The HMM mixture produces an activity distribution over HERVs. The reliability of the distribution can be estimated with a bootstrap-like method as follows: The EST data are resampled with replacement several times and then the activities are reoptimized for each replicate while other parameters are kept fixed (see Publication 6 for details). A similar approach can also be used to estimate the reliability of the activity distribution obtained with the BLAST approach; the EST counts are recomputed for each replicate.

6.3 Experiments

6.3.1 Validation with simulated data

The performance of the HMM mixture was evaluated using a simulated data set in Publication 6. The mixture model was also compared to the heuristic BLAST alternative.

A set of artificial ESTs was generated from a representative set of 181 HERVs (see Section 6.1.1) using the HMM mixture. To make the simulated EST data realistic, the parameters of the generating HMM were set close to the parameters learned from real ESTs, and the lengths of the ESTs were controlled by rejecting too short and too long ESTs. After generation the ESTs were processed exactly the same way as real data, starting with BLAST to match the HERVs against the ESTs.

Both the HMM model and the heuristic BLAST approach were applied to the simulated data set. The relative activity distributions learned by these two models were compared to the true activity distribution (the one used while generating the artificial ESTs). The results from this comparison are shown in Figure 6.5. As can be seen, both the HMM model and the BLAST approach closely follow the true activity distribution.

The performance of the two approaches was also quantified using the Kullback-Leibler divergence to measure the distance between the true and the learned distribution. In this comparison the HMM method performs slightly better than the BLAST approach (for details see Publication 6). The surprisingly good performance of the BLAST approach suggests that it can be used for large-scale studies where HMM training would be computationally too costly.

The difference between the two approaches was the most notable in the case of the HML2 group, which is a young HERV group containing almost identical sequences. For the HML2 group, and for other young families, it is preferable to use the rigorous probabilistic approach, i.e., the HMM mixture. The mixture model was used to study HML2 HERVs in Publication 7.

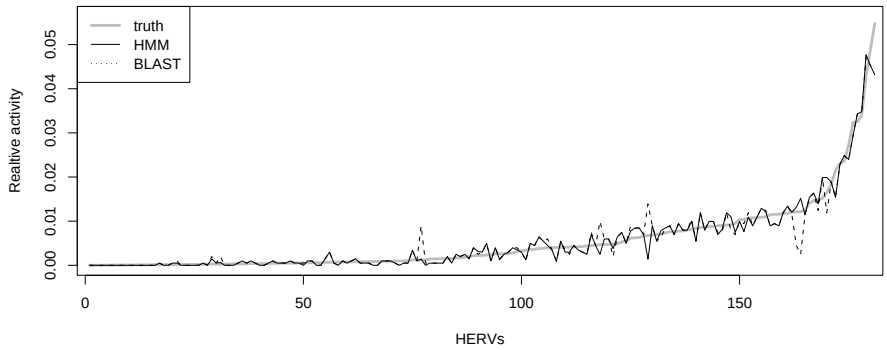


Figure 6.5: Activities of selected HERVs (simulated data). HMM model is able to estimate the true underlying activities slightly better than the heuristic BLAST approach. The figure is taken from Publication 6.

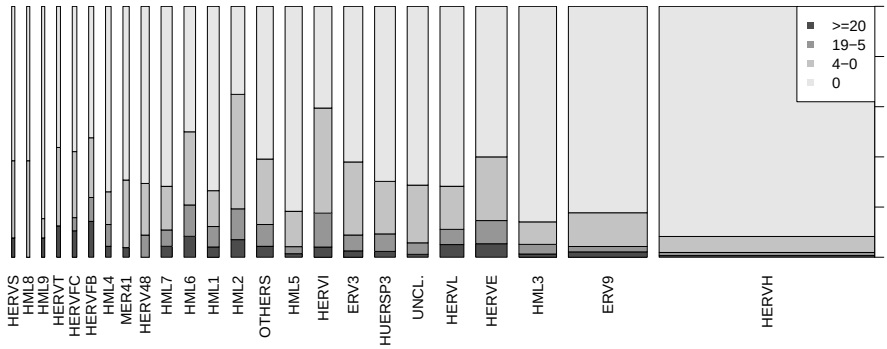


Figure 6.6: The proportion of active and inactive HERVs varies a lot from group to group. The proportion of active HERVs in each group is presented with a stacked area plot. The two darkest gray areas together show the proportion of active HERVs in that group, the lightest gray area shows the proportion of inactive HERVs. The area in between shows HERVs with 1-4 EST hits; for these HERVs the activity status is uncertain (see Publication 6 for more details). The widths of the bars are proportional to the size of the HERV group. The figure is taken from Publication 6.

6.3.2 Overview of HERV activation

In Publication 6 the activities of a large set of 2450 HERVs were explored to get an overview of HERV expression in humans and to detect individual active HERVs. To save time the activities were estimated using the fast BLAST-based approach.

Here the main biological results from Publication 6 are briefly reviewed. First, about 7% of the HERVs were active (had at least 5 EST hits in their gene or LTR areas) and most of the rest were completely inactive based on the EST collection used; 1903 HERVs had no EST matches. Second, almost all groups have some active elements. However, the proportion of active HERVs varies considerably from group to group (see Fig. 6.6). Third, the observed expression had many forms. The expectation that only young, intact elements would be able to activate proved to be wrong as there were several kinds of HERVs among the active ones: old and young, full-length and those missing several viral genes, HERVs with open reading frames in their genes and HERVs that can not produce viral proteins.

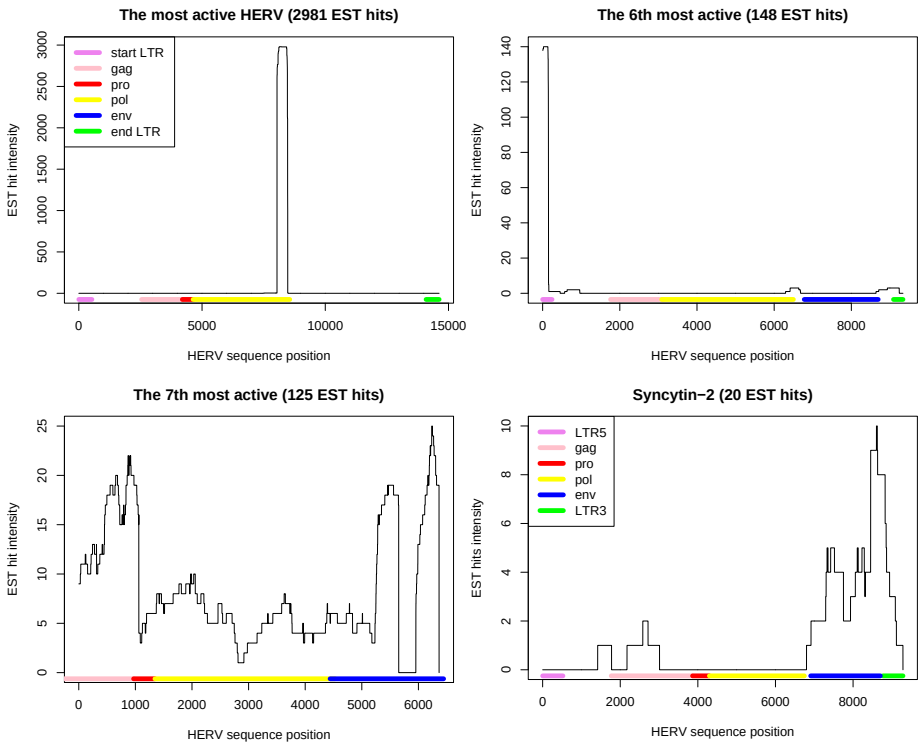


Figure 6.7: EST hit locations for a few of the active HERVs. The colored blocks below the curve represent the HERV structure (genes, LTRs) and the curve presents EST hit intensity along the HERV structure. The HERVs shown on the top row exhibit fragmented expression. It is highly unlikely that these HERVs produce retroviral proteins. However, the HERV in the top right corner is presumably functioning as an (alternative) end for a nearby gene. In contrast, the HERVs shown on the bottom row may be coding for retroviral proteins. Actually, the HERV on the right contains the syncytin-2 gene, a known human gene originating from the env-gene of this HERV locus. The HERV on the left may also be retrovirally active. The figure is taken from Publication 6.

Fourth, the observed expression patterns were surprising. The EST sequences do not necessarily match the HERV in the retroviral gene areas, but exhibit fragmented expression. This suggests that in many cases the retroviral sequence has been used as a building block for something else than retroviral proteins, for example human gene exons, promoters, or polyadenylation signals. The evidence for this is (i) only a few of the active elements have viral gene open reading frames, and (ii) the ESTs often match only a short portion of a viral gene (see the top row of Fig. 6.7 for examples). However, for some HERVs there may be an alternative explanation: the retroviral transcripts may have RNA-mediated activities.

6.3.3 Expression profiles of HML2 HERVs

In Publication 7 expression profiles of individual HERV sequences across a set of tissues were studied. The expression profile enables the study of the differential expression patterns of individual HERVs, leading to a better understanding of the function of individual HERVs. For example, HERVs that are more/only active

in the brain tissue may have functions related to neurodegenerative diseases or to normal brain functions. The profiling approach is widely used in the study of human gene function, see Chapter 4.

The HML2 group was selected for analysis because it has the largest proportion of relatively intact elements and because it was found to contain the largest percentage of active elements (see Publication 6 and Fig. 6.6). Some of the HML2 sequences are full-length, i.e. have retained the typical retrovirus structure "LTR-gag-pro-pol-env-LTR", and few of these even have open reading frames for the env gene, i.e. they could produce retroviral env proteins.

In addition to the HML2 HERVs, some extra HERVs were added to the HERV set in order to capture cross-talk between the HERV groups. The extra HERVs were selected broadly from all HERV groups. However, only HERVs that were estimated to be active by the heuristic BLAST approach were included (see section 6.2.2 and Publication 7).

The HMM model was able to estimate the activity profiles of the HML2 HERVs reliably, according to the bootstrap-based reliability estimation (Section 6.2.4). The activity profiles are shown in Figure 6.8A. Many of the HERVs exhibit tissue-specific expression. There are also some HERVs that are active in all tissues, as well as HERVs that are not active in any of them. The activities of most HML2 HERVs were previously unknown. Publication 7 gives a more detailed analysis of sample HERVs.

The results show that several of the extra HERVs (non-HML2 HERVs) are very active (Fig. 6.8B). Furthermore, in each case the probability mass allotted to them was more than half of the total (ranging from 63% in the placenta to 52% in the lungs). This verifies the assumption that there is cross-talk between families. ESTs retrieved using the HML2 HERVs as queries actually match the extra HERVs better. To conclude, it was necessary to add the extra HERVs to get reliable activity estimates for HML2 HERVs.

6.4 Discussion and Conclusions

In this chapter a hidden Markov mixture model developed for the task of estimating HERV activities from EST sequences was introduced. An extended version of the model can be used to estimate expression profiles of HERVs over a set of conditions. The results obtained with the HMM mixture were found to be reliable according to a bootstrap-based reliability estimation.

The HMM model was compared to a heuristic BLAST-based alternative with experiments on simulated data. Both methods were able to estimate underlying activities fairly well. The surprisingly good performance of the computationally simpler alternative justifies its use when the rigorous probabilistic method would be too slow. It is still recommended that the more accurate HMM model is used in smaller-scale studies, in particular for the more difficult HERV groups (groups containing close to identical sequences).

In Publication 6 the activities of all individual HERV sequences were estimated in order to gain an overall picture of HERV activity in humans. The results are biologically interesting and merit further study. The individual HERVs reported as active with our method can later be verified with laboratory methods; by contrast, exhaustive search of active HERVs with laboratory methods would be too expensive.

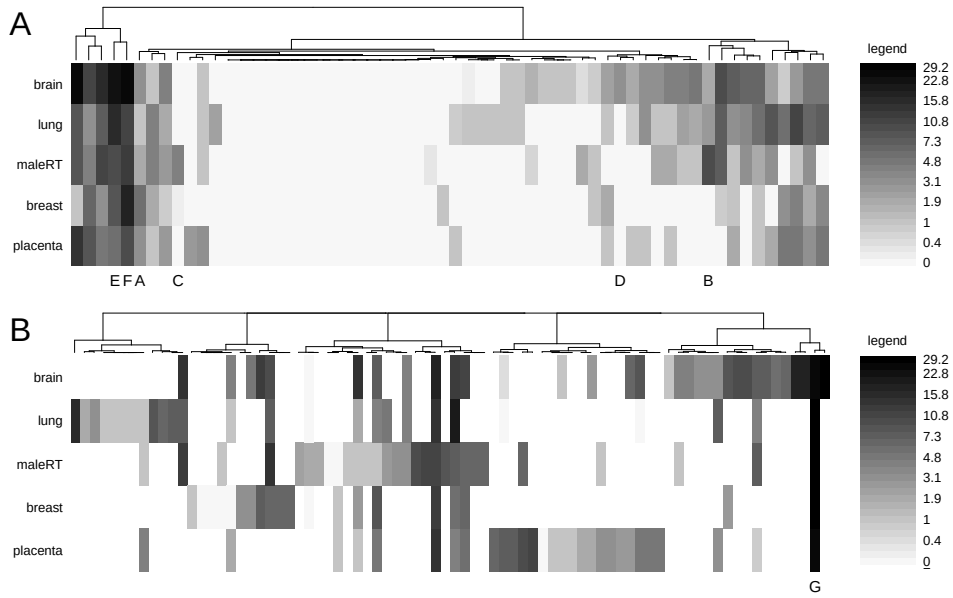


Figure 6.8: The activities of the HML2 (panel **A**) and non-HML2 (panel **B**) HERVs. In both panels the rows depict the activity distributions over HERVs and the columns the expression profiles of individual HERV sequences. The activity values are shown on a logarithmic scale, as can be seen from the legends on the right. The scale is the same in both panels. The numbers next to the legend are the probabilistic EST counts for each gray shade. The columns have been ordered according to a hierarchical clustering based on the (unlogarithmic) Euclidean distances between the HERV expression profiles. The figure is taken from Publication 7.

In Publications 6 and 7 the quality of the HERV data caused some problems for the analysis. The aim was to study the expression of all HERVs, not only the full-length elements but also the fragmented HERVs that are harder to detect from the human genome. In the process of ensuring that the more mutated HERVs are not missed some elements that are combinations of retrovirus and retrotransposon sequences may have been included into the HERV set produced by RetroTector. The problems caused by such chimeric elements, and elements containing long portions of DNA not annotated as a viral gene or LTR, were handled differently in the two publications. In Publication 6 all HERVs where the activity was in un-annotated portions were removed and the analysis was focused on retrovirus originating activity. In Publication 7, however, the HERVs were included as is. Results from the latter publication indicate that some of the activity in the un-annotated areas is due to active L1 retrotransposons and SVA elements (see the publication for details).

The proposed HMM method is generally applicable; it can be used to study endogenous retroviruses in other organisms, or to include other kinds of transposable elements. The source of active sequences could also be something else than ESTs. For example, the method could be used as a post-processing step in an RT-PCR reaction where a broadly targeting primer (all members of a HERV group are amplified) is used. When the PCR products are sequenced, they can be compared to the members of the targeted HERV group using the HMM mixture.

Chapter 7

Conclusion

In this thesis self-organizing map (SOM)-based exploratory data analysis (EDA) approaches were developed and applied successfully to genomic data sets. The cluster displays constructed with SOMs show similar clusterings as displays formed by hierarchical clustering (HC) or phylogenetic trees (PT), supporting the validity of the SOM approach. The advantage of SOM is the two dimensional visualization display that shows the relationships of the clusters as well as similarities between individual samples. The SOM presentation allows intuitive exploration of the data.

In the first part of the thesis (Publications 1-2) the SOM was applied to the analysis of a yeast gene expression data set. It was demonstrated that SOMs can find meaningful biological information from gene expression data. Naturally, the SOM approach can be used to analyze gene expression data from other organisms in an analogous manner.

In the second part of the thesis (Publications 3-5) the SOM was applied to the study of human endogenous retroviruses (HERVs) and their classification. The results showed that the SOM was able to extract new knowledge from a HERV sequence collection, subsets of which had previously been analyzed with PTs. The PTs and the SOM can complement each other when constructing a final grouping for all HERV sequences. The PTs represent the evolutionary connections between groups of sequences, but the rigorous PT inference algorithms are, unfortunately, not able to handle very large data sets. The SOM, on the other hand, is well suited for analyzing larger collections of sequences simultaneously. The SOM approach can be applied to the study of other kinds of (biological) sequences as well. One potential application area are members of protein families from several organisms.

The choice of metric directly affects EDA results. The metric should be chosen such that the ensuing similarities will be informative for the analysis goal. In this thesis, the performance of different distance measures in their ability to represent the functional classification of genes was compared for a few data sets. The correlation coefficient was the best suggesting that only the relative values of the expression ratios are important. The functional classification of genes can also be built directly into the distance measure using the learning metrics principle. The new location-specific metric measures changes in gene expression but weights the changes according to how much they contribute to changes in the functional classes. The learning metrics offer a way to incorporate prior biological knowledge about the function of the genes into the measurements.

The learning metrics principle is general and can be used with various types

of vectorial data sets. Furthermore, the auxiliary data can be also something else than a pre-existing classification. For example, in gene expression data analysis the auxiliary data may be information about gene expression levels of homologous genes in another organism. This way information from model organisms, that have been studied extensively, can be utilized when analyzing the genes of a less studied organism, like human. Besides SOMs the learning metrics principle can be used in conjuction with various methods, such as Sammon's mapping (Peltonen et al., 2004), discriminant analysis (Peltonen and Kaski, 2005) and clustering (Sinkkonen and Kaski, 2002; Kaski et al., 2005a,b). Extension of learning metrics for non-metric data sources, like sequences and other structured data, would enable it to be used in conjunction with the Median SOM and other methods where the input data is represented with a pairwise similarity data matrix. This is an idea for future work.

The visualization aspect of the SOM was compared to other unsupervised methods using the new trustworthiness measure that gives a good score for visualization displays with few false positive proximities. For gene expression data sets the SOM was found to be the most trustworthy alternative. On the other hand, the results obtained from the comparison of PT and median SOM suggest that the performance of the median SOM is not necessarily better than some of the alternatives. A more extensive comparison is needed to clarify the differences between the median SOM and other visualization and clustering methods for pairwise data sets.

Results from Publications 4 and 5 demonstrate that visualization of the reliability of different parts of the SOM display is a valuable help in SOM-based data analysis. In this thesis two new measures were introduced for estimating the reliability of each map unit on a SOM display, a bootstrap-based and a trustworthiness-based measure. Of these two measures the trustworthiness one is preferred. The reliability visualization, where each map unit is colored according to its reliability, is used to help focus the analysis of the SOM to the most reliable areas of the map. The reliability estimates can also be applied in other cases where a SOM is used. The estimates are not limited to pairwise distance matrices, but can be applied also to vectorial SOMs. Furthermore, the trustworthiness based reliability estimate can be applied to various other visualization methods. For example, all individual points in a projection or all leaves and inner vertices in a hierarchical clustering tree could be colored according to their trustworthiness.

The EDA process usually continues with a closer analysis of interesting groups of samples selected manually based on the SOM display, a reliability visualization, and all available background information. In this thesis bootstrap-based reliability estimates for validating the compactness and purity of such groups were presented. Before the group is analyzed closer it should be verified that the group forms a true cluster in the data. This can be done using the new measures. Again, the measures can be applied in all types of SOM applications.

In the last part of the thesis (Publications 6-7) a computational method for estimating HERV activities was developed. The generative hidden Markov mixture (HMM) model estimates the activities of individual HERVs rather than those of HERV groups, i.e., is able to overcome the limitations of commonly used laboratory techniques. HERVs reported as active using the HMM mixture can later be verified with detailed laboratory methods; by contrast, exhaustive search of active individual HERVs with the difficult laboratory procedures would be too expensive. The computational approach allows exploration for potentially active HERVs. Us-

ing the HMM method a more detailed picture of HERV activity in real data was obtained. Many of the active HERVs exhibit fragmented activity patterns and are likely to serve purposes other than the production of retroviral proteins. For example, they may be human gene exons or promoters, or have an RNA-mediated function. These as well as the few potentially retrovirally active HERVs that were detected should be analyzed more closely in future studies.

The proposed HMM method is generally applicable; it can be used to study endogenous retroviruses in other organisms, or to include other kinds of transposable elements. The source of active sequences could also be something else than ESTs. The EST matching problem for which the model has been designed is similar to the cross-hybridization problem in tiling microarrays. A future research direction could be to extend the HMM model for the microarray cross-hybridization problem.

The new HMM model can be used for estimating expression profiles of HERVs. In future studies the HERV expression data set could then be explored using the SOM-based methods introduced in the first part of this thesis. The problem of understanding the control mechanisms behind HERV expression is also highly relevant. Furthermore, the learning metrics principle can be applied to HERVs, for example, in conjuction with human gene expression data. The HERV sequences may control the expression of nearby genes. Then it would make sense to guide the gene expression data analysis with auxiliary information obtained from HERVs located near to the gene in the DNA.

Bibliography

T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura. Informatics for unveiling hidden genome signatures. *Genome Reseach*, 13(4):693–702, 2003.

T. Abe, H. Sugawara, S. Kanaya, M. Kinouchi, and T. Ikemura. Self-organizing map (SOM) unveils and visualizes hidden sequence characteristics of a wide range of eukaryote genomes. *Gene*, 365:27–34, 2006.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

A.-C. Andersson, Z. Yun, G. O. Sperber, E. Larsson, and J. Blomberg. ERV3 and related sequences in humans: Structure and RNA expression. *Journal of Virology*, 79 (4):9270–9284, 2005.

A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc. Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, 19(5):563–570, 2003.

Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17:S22–S29, 2001.

Z. Bar-Joseph, E. Demaine, D. K. Gifford, N. Srebro, A. Hameland, and T. S. Jaakkola. K-ary clustering with optimal leaf ordering for gene expression data. *Bioinformatics*, 19:1070–8, 2003.

A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Research*, 30(1):276–280, 2002.

L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

S. Becker and M. Plumbley. Unsupervised neural network learning procedures for feature extraction and classification. *Journal of Applied Intelligence*, 6:1–21, 1996.

M. Beer and S. Tavazoie. Predicting gene expression from sequence. *Cell*, 117:185–198, 2004.

M. Belkin and P. Niyogi. Laplacian Eigenmaps and spectral techniques for embedding and clustering. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 585–591, Cambridge, MA, 2002. MIT Press.

S. Ben-David, U. von Luxburg, and D. Pál. A sober look at clustering stability. In *19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006.*, volume 4005/2006 of *Lecture Notes in Computer Science*, pages 5–19, 2006.

A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *Pasific Symposium on Biocomputing*, volume 7, pages 6–17, 2002.

A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98:13790–13795, 2001.

S. Blaise, N. de Parseval, L. Bénit, and T. Heidmann. Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. *Proceedings of the National Academy of Sciences*, 100(22):13013–13018, 2003.

J. Blomberg, D. Uschameckis, and P. Jern. Evolutionary aspects of human endogenous retroviral sequences and disease. In E. Sverdlov, editor, *Retroviruses and Primate Evolution*, pages 208–243. Eurekah Bioscience, 2005.

I. Borg and P. Groenen. *Modern Multidimensional Scaling*. Springer, New York, 1997.

E. Bosch and M. A. Jobling. Duplications of the AZFa region of the human Y chromosome are mediated by homologous recombination between HERVs and are compatible with male fertility. *Human Molecular Genetics*, 12(3):341–347, 2003.

R. J. Britten. Mobile elements inserted in the distant past have taken on important functions. *Gene*, 205:177–182, 1997.

J. Bryan. Problems in gene clustering based on gene expression data. *Journal of Multivariate Analysis*, 90:44–66, 2004.

D. Butler. Are you ready for the revolution? *Nature*, 409:758–760, 2001.

T. Christensen, P. D. Sorensen, H. J. Hansen, and A. Moller-Larsen. Antibodies against a human endogenous retrovirus and the preponderance of env splice variants in multiple sclerosis patients. *Multiple Sclerosis*, 9(1):6–15, 2003.

J. M. Coffin, S. H. Hughes, and H. E. Varmus. *Retroviruses*. Cold Spring Harbor Laboratory Press, 1997. Available at: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=rv.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

E. de Bodt and M. Cottrell. Bootstrapping self-organising maps to assess the statistical significance of local proximity. In *Proceedings of ESANN'2000, 8th European Symposium on Artificial Neural Networks.*, pages 245–254, 2000.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

S. Depil, C. Roche, P. Dussart, and L. Prin. Expression of human endogenous retrovirus, HERV-K, in the blood cells of leukemia patients. *Leukemia*, 16(2):254–259, 2002.

S. Drăghici. *Data analysis tools for DNA microarrays*. Chapman & Hall / CRC, 2003.

B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1): 1–26, 1979.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95:14863–14868, 1998.

J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

E. A. Ferrán and P. Ferrara. Topological maps of protein sequences. *Biological cybernetics*, 65(6):451–458, 1991.

W. M. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416, 1971.

G. Forney. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

A. Forsman, Z. Yun, D. Uzhameckis, P. Jern, L. Hu, and J. Blomberg. Development of broadly targeted human endogenous retroviral pol -based real time -PCRs. Quantitation of RNA expression in human tissues. *Journal of Virological Methods*, 129:16–30, 2005.

O. Frank, M. Giehl, C. Zheng, R. Hehlmann, C. Leib-Mösch, and W. Seifarth. Human endogenous retrovirus expression profiles in samples from brains of patients with schizophrenia and bipolar disorders. *Journal of Virology*, 79(17):10890–10901, 2005.

G. Giaver et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418(387–91), 2002.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis (2nd edition)*. Chapman & Hall/CRC, 2003.

F. D. Gibbons and F. P. Roth. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Research*, 12:1574–1581, 2002.

R. Gifford and M. Tristem. The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes*, 26(3):291–315, 2003.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

G. J. Goodhill and T. J. Sejnowski. A unifying objective function for topographic mappings. *Neural Computation*, 9(6):1291–1303, 1997.

D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, Cambridge, MA, 2001.

J. Handl, J. Knowles, and D. B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005.

D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18:S145–S154, 2002.

J. Hanke and J. G. Reich. Kohonen map as visualization tool for the analysis of protein sequences: Multiple alignments, domains and segments of secondary structures. *Bioinformatics*, 12(6):447–454, 1996.

T. Hastie, R. Tibshirani, and A. Buja. Flexible discriminant and mixture models. In J. Kay and D. Titterington, editors, *Neural Networks and Statistics*. Oxford University Press, 1995.

S. Hautaniemi, O. Yli-harja, J. Astola, P. Kauraniemi, A. Kallioniemi, M. Wolf, J. Ruiz, S. Mousses, and O.-P. Kallioniemi. Analysis and visualization of gene expression microarray data in human cancer using self-organizing maps. *Machine Learning*, 52: 45–66, 2003.

G. Hinton and S. Roweis. Stochastic neighbor embedding. In S. T. S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 833–840. MIT Press, 2002.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–41,498–520, 1933.

L. Hu, D. Hornung, R. Kurek, H. Östman, J. Blomberg, and A. Bergqvist. Expression of human endogenous retroviruses in endometriosis and ovarian cancer. *AIDS Research and Human Retroviruses*, 22:551–557, 2006.

J. C. Huang, A. Kannan, and J. Winn. Bayesian association of haplotypes and non-genetic factors to regulatory and phenotypic variation in human populations. *Bioinformatics*, 23(13), 2007.

T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffrey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

R. Hughey and A. Krogh. SAM: Sequence alignment and modeling software system. Technical Report UCSC-CRL-95-7, University of California, Santa Cruz, CA, 1995.

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

T. Jaakkola, M. Diekhans, and D. Haussler. Using the Fisher kernel method to detect remote protein homologies. In *Proceedings of the 7th Annual International Conference on Intellegent Systems for Molecular Biology (ISMB'99)*, pages 1149–1158. AAAI Press, 1999.

A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

A. K. Jain and J. V. Moreau. Bootstrap technique in cluster analysis. *Pattern Recognition*, 20(5):547–568, 1987.

P. Jern, G. O. Sperber, and J. Blomberg. Definition and variation of human endogneous retrovirus H. *Virology*, 327:93–110, 2004.

I. Jordan, I. B. Rogozin, G. V. Galzko, and E. V. Koonin. Origin of substantial fraction of human regulatory sequences from transposable elements. *Trends in Gentics*, 19(2): 68–72, 2003.

J. Jurka. Repbase update: A database and an electronic journal of repetitive elements. *Trends in Genetics*, 16:418–420, 2000.

A.-K. Järvinen, S. Hautaniemi, H. Edgren, P. Auvinen, J. Saarela, O.-P. Kallioniemi, and O. Monni. Are data from different gene expression microarray platforms comparable? *Genomics*, 83(6):1164–1168, 2004.

S. Kanaya, M. Kinouchi, T. Abe, Y. Kudo, Y. Yamada, T. Nishi, H. Mori, and T. Ikemura. Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the E. coli o157 genome. *Gene*, 276(1–2):89–99, 2001.

M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resources for deciphering the genome. *Nucleic Acids Research*, 32:D270–D280, 2004.

S. Kaski and J. Peltonen. Informative discriminant analysis. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, pages 329–336. AAAI Press, 2003.

S. Kaski and J. Sinkkonen. Metrics that learn relevance. In *Proceedings of IJCNN-2000, International Joint Conference on Neural Networks*, volume V, pages 547–552. IEEE Service Center, 2000.

S. Kaski and J. Sinkkonen. Principle of learning metrics for exploratory data analysis. *Journal of VLSI Signal Processing, special issue on Machine Learning for Signal Processing*, 37:177–188, 2004.

S. Kaski, J. Kangas, and T. Kohonen. Bibliography of self-organizing map (SOM) papers: 1981–1997. *Neural Computing Surveys*, 1(3&4):1–176, 1998.

S. Kaski, J. Venna, and T. Kohonen. Coloring that reveals high-dimensional structures in data. In T. Gedeon, P. Wong, S. Halgamuge, N. Kasabov, D. Nauck, and K. Fukushima, editors, *Proceedings of ICONIP'99, 6th International Conference on Neural Information Processing*, volume II, pages 729–734. IEEE Service Center, 1999.

S. Kaski, J. Sinkkonen, and J. Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12:936–947, 2001.

S. Kaski, J. Nikkilä, J. Sinkkonen, L. Lahti, J. Knuuttila, and C. Roos. Associative clustering for exploring dependencies between functional genomics data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, Special Issue on Machine Learning for Bioinformatics – Part 2*, 2(3):203–216, 2005a.

S. Kaski, J. Sinkkonen, and A. Klami. Discriminative clustering. *Neurocomputing*, 69: 18–41, 2005b.

T.-H. Kim, Y.-J. Jeon, W.-Y. Kim, and H.-S. Kim. HESAS: HERVs expression and structure analysis system. *Bioinformatics*, 21(8):1699–1700, 2005. doi: 10.1093/bioinformatics/bti194.

T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

T. Kohonen. *Self-Organizing Maps*. Springer, 3rd edition, 2001.

T. Kohonen. Self-organizing maps of symbol strings. Technical Report A42, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.

T. Kohonen and P. Somervuo. How to make large self-organizing maps for nonvectorial data. *Neural Networks*, 15:945–952, 2002.

T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela. Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11:574–585, 2000.

A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler. Hidden Markov models in computational biology. applications to protein modeling. *Journal of Molecular Biology*, 235(5):1501–1531, 1994.

J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrica*, 29(1–27), 1964.

R. Kustra and A. Zagdanski. Incorporating gene ontology in clustering gene expression data. In *CBMS'06: Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems*, pages 555–563. IEEE Computer Society, 2006.

K. Lagus, S. Kaski, and T. Kohonen. Mining massive document collections by the WEBSOM method. *Information Sciences*, 163:135–156, 2004.

G. R. Lanckriet, T. D. Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

E. S. Lander. Array of hope. *Nature Genetics*, 21:3–4, 1999.

C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble. Mismatch string kernels for discriminative protein classification. *Bioinormatics*, 20(4):467–76, 2004.

E. Levine and E. Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13(11):2573–2593, 2001.

D. Lockhart, H. Dong, M. Byrne, M. Follettie, M. Gallo, M. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–80, 1996.

D. J. Lockhart and E. A. Winzeler. Genomics, gene expression and DNA arrays. *Nature*, 405:827–836, 2000.

R. Löwer. The pathogenic potential of endogenous retroviruses: facts and fantasies. *Trends in Microbiology*, 7(9):350–356, 1999.

G. Lunter. Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics*, 23(13):i289–i296, 2007.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297. University of California Press, 1967.

S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1: 24–45, 2004.

D. L. Mager and P. Medstrand. Retroviral repeat sequences. In *Encyclopedia of the Human Genome*, pages 57–63. Nature Publishing Group, 2003.

S. Mahony, J. O. McInerney, T. J. Smith, and A. Golden. Gene prediction using the self-organizing map: automatic generation of multiple gene models. *BMC Bioinformatics*, 5:23, 2004.

S. Mahony, A. Golden, T. J. Smith, and P. V. Benos. Improved detection of DNA motifs using a self-organized clustering of familial binding profiles. *Bioinformatics*, 21(Suppl. 1):i283–i291, 2005a.

S. Mahony, D. Hendrix, A. Golden, T. J. Smith, and D. S. Rokhsar. Transcription factor binding site identification using the self-organizing map. *Bioinformatics*, 21(9): 1807–14, 2005b.

S. Mahony, P. V. Benos, T. J. Smith, and A. Golden. Self-organizing neural networks to support the discovery of DNA-binding motifs. *Neural Networks*, 19:950–962, 2006.

P. Mangiameli, S. K. Chen, and D. West. A comparison of SOM neural network and hierarchical clustering methods. *European Journal of Operational Research*, 93:402–17, 1996.

J. Mao and A. Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks*, 6(2):296–317, 1995.

K. V. Mardia. Statistics of directional data. *Journal of the Royal Statistical Society. Series B*, 37:349–393, 1975.

J. Mayer and E. U. Meese. The human endogenous retrovirus family HERV-K(HML-3). *Genomics*, 80(3):331–343, 2002.

G. J. McLachlan and K. E. Basford. *Mixture Models. Inference and Applications to Clustering*. Marcel Dekker, 1988.

G. J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

L. M. McShane, M. D. Radmacher, B. Freidlin, R. Yu, M.-C. Li, and R. Simon. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18(10):1462–1469, 2002.

P. Medstrand, L. van de Lagemaat, C. Dunn, J.-R. Landry, D. Svenback, and D. Mager. Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenetic and Genome Research*, 110(1-4):342–352, 2005.

S. Mi, X. Lee, X.-P. Li, G. M. Veldman, H. Finnerty, L. Racie, E. LaVallie, X.-Y. Tang, P. Edouard, S. Howes, J. C. Keith Jr., and J. M. McCoy. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, 403: 785–789, 2000.

S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1-2):91–118, 2003.

A. Muir, A. Lever, and A. Moffett. Expression and functions of human endogenous retroviruses in the placenta: An update. *Placenta*, 25(Suppl. 1):S16–S25, 2004.

S. Muradrasoli, A. Forsman, L. Hu, V. Blikstad, and J. Blomberg. Development of real-time pcrs for detection and quantitation of human MMTV-like (HML) sequences. HML expression in human tissues and cell lines. *Journal of Virological Methods*, 136:83–92, 2006.

Nature Genetics Supplement. The chipping forecast. *Nature Genetics*, 21(1s):1–60, 1999.

P. N. Nelson, P. R. Carnegie, J. Martin, H. Davari Ejtehadi, P. Hooley, D. Roden, S. Rowland-Jones, P. Warren, J. Astley, and P. G. Murray. Demystified ... human endogenous retroviruses. *Molecular Pathology*, 56:11–18, 2003.

J. Nikkilä, P. Törönen, S. Kaski, J. Venna, E. Castrén, and G. Wong. Analysis and visualization of gene expression data using self-organizing maps. *Neural Networks*, 15 (8-9):953–966, 2002.

M. Oja, S. Kaski, and T. Kohonen. Bibliography of self-organizing map (SOM) papers: 1998-2001 addendum. *Neural Computing Surveys*, 3:1–156, 2003.

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

S. Patzke, M. Lindeskog, E. Munthe, and H. C. Aasheim. Characterization of a novel human endogenous retrovirus, HERV-H/F, expressed in human leukemia cell lines. *Virology*, 303(1):164–173, 2002.

W. Pearson and D. Lipman. Improved tools for biological sequence comparision. *Proceedings of the National Academy of Sciences*, 85:2444–2448, 1988.

J. Peltonen. Data exploration with learning metrics. *Dissertations in Computer and Information Science*, report D7, 2004. PhD Thesis, Helsinki University of Technology, Finland.

J. Peltonen and S. Kaski. Discriminative components of data. *IEEE Transactions on Neural Networks*, 16:68–83, 2005.

J. Peltonen, A. Klami, and S. Kaski. Learning more accurate metrics for self-organizing maps. In J. R. Dorronsoro, editor, *Artificial Neural Networks—ICANN 2002*, pages 999–1004. Springer, 2002.

J. Peltonen, A. Klami, and S. Kaski. Improved learning of Riemannian metrics for exploratory analysis. *Neural Networks*, 17:1087–1100, 2004.

M. Pöllä, T. Honkela, and T. Kohonen. Bibliography of self-organizing map (SOM) papers: 2002-2005 addendum. *Neural Computing Surveys*, 2008. Forthcoming.

J. L. Portis. Perspectives on the role of endogenous human retroviruses in autoimmune diseases. *Virology*, 296:1–5, 2002.

J. Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–27, 2001.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

D. Rogers and T. Tanimoto. A computer program for classifying plants. *Science*, 132 (3434):1115–1118, 1960.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

J. W. Sammon Jr. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18:401–409, 1969.

M. Schena, D. Shalon, R. Davis, and P. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

A. Schliep, I. G. Costa, C. Steinhoff, and A. Schönhuth. Analyzing gene expression time-courses. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2 (3):179–93, 2005.

M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

E. Segal and D. Koller. Probabilistic hierarchical clustering for biological data. In *Proceedings of the 6th international conference on Research in Computational Molecular Biology (RECOMB)*, pages 273–280, 2002.

W. Seifarth, B. Spiess, U. Zeifelder, C. Speth, R. Hehlmann, and C. Leib-Mösch. Assessment of retroviral activity using a universal retrovirus chip. *Journal of Virological Methods*, 112:89–91, 2003.

W. Seifarth, O. Frank, U. Zeifelder, B. Spiess, A. D. Greenwood, R. Hehlmann, and C. Leib-Mösch. Comprehensive analysis of human endogenous retrovirus transcriptional activity in human tissues with a retrovirus-specific microarray. *Journal of Virology*, 79(1):341–52, 2005.

A. Sesto, M. Navarro, F. Burslem, and J. L. Jorcano. Analysis of the ultraviolet B response in primary human keratinocytes using oligonucleotide microarrays. *Proceedings of the National Academy of Sciences*, 99(5):2965–70, 2002.

S. P. Shah, W. L. Lam, R. T. Ng, and K. P. Murphy. Modeling recurrent DNA copy number alterations in array cgh data. *Bioinformatics*, 23(13):i450–i458, 2007.

B. S. Shastry. Schizophrenia: a genetic perspective (review). *International Journal of Molecular Medicine*, 9(3):207–212, 2002.

M. Shiga, I. Takigawa, and H. Mamitsuka. Annotating gene function by combining expression data with a modular gene network. *Bioinformatics*, 23:i468–i478, 2007.

B. W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, 1986.

J. Sinkkonen. Learning metrics and discriminative clustering. *Dissertations in Computer and Information Science*, report D2, 2003. PhD Thesis, Helsinki University of Technology, Finland.

J. Sinkkonen and S. Kaski. Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, 14:217–239, 2002.

A. F. A. Smit, R. Hubley, and P. Green. Repeatmasker open-3.0., 1996-2004. http://www.repeatmasker.org.

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

K.-A. Sohn and E. P. Xing. Spectrum: joint bayesian inference of population structure and recombination events. *Bioinformatics*, 23(13):i490–i498, 2007.

G. Sperber, P. Jern, T. Airola, and J. Blomberg. Automated recognition of retroviral sequences – RetroTector©. *Nucleic Acids Research*, 35(15):4964–76, 2007.

Y. Stauffer, G. Theiler, P. Sperisen, Y. Lebedev, and C. V. Jongeneel. Digital expression profiles of human endogenous retroviral families in normal and cancerous tissues. *Cancer Immunity*, 4:2, 2004.

A. I. Su, M. P. Cooke, K. A. Ching, Y. Hakak, J. R. Walker, T. Wiltshire, A. P. Orth, R. G. Vega, L. M. Sapinoso, A. Moqrich, A. Patapoutian, G. M. Hampton, P. G. Schultz, and J. B. Hogenesch. Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences*, 99:4465–4470, 2002.

P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrowsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, 96:2907–2912, 1999.

A. Tanay, R. Sharan, and R. Shamir. *Handbook of Computational Molecular Biology*, chapter Biclustering algorithms: A survey, pages 26–1—26–17. Chapman and Hall/CRC Press, 2006.

S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22(3):281–285, 1999.

J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, 1994.

P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén. Analysis of gene expression data using self-organizing maps. *FEBS Letters*, 451:142–146, 1999.

K. Toutanova, F. Chen, K. Popat, and T. Hofmann. Text classification in a hierarchical mixture model for small training sets. In *Proceedings 10th International Conference on Information and Knowledge Management*, pages 105 – 113, 2001.

Y. Tu, G. Stolovitzky, and U. Klein. Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences*, 99(22): 14031–14036, 2002.

E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983.

J. Tukey. *Exploratory data analysis*. Addison-Wesley, 1977.

A. Ultsch and H. Siemon. Kohonen's self organizing feature maps for exploratory data analysis. In *Proc. INNC'90, Int. Neural Network Conf.*, pages 305–308. Kluwer, 1990.

J. Venna. Dimensionality reduction for visual exploration of similarity structures. *Dissertations in Computer and Information Science*, report D20, 2007. PhD Thesis, Helsinki University of Technology, Finland.

J. Venna and S. Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Proceedings of the International Conference on Artificial Neural Networks—ICANN 2001; Vienna.*, pages 485–491. Berlin: Springer, 2001.

J. Venna and S. Kaski. Local multidimensional scaling. *Neural Networks*, 19:889–99, 2006.

J. Venna and S. Kaski. Comparison of visualization methods for an atlas of gene expression data sets. *Information Visualization*, 6:139–54, 2007a.

J. Venna and S. Kaski. Nonlinear dimensionality reduction as information retrieval. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS*07), San Juan, Puerto Rico, March 21-24*, 2007b.

J. Vesanto and E. Alhoniemi. Clustering of self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, 2000.

J. Vilo, A. Brazma, I. Jonassen, A. Robinson, and E. Ukkonen. Mining for putative regulatory elements in the yeast genome using gene expression data. In *Proc. of Intelligent Systems for Molecular Biology*, pages 382–394, 2000.

A. Vinokourov and M. Girolami. A probabilistic hierarchical clustering method for organizing collections of text documents. In *Proc. of 15th International Conference on Pattern Recognition (ICPR'00)*, volume 2, pages 182–185, 2000.

A. Vinokourov and M. Girolami. A probabilistic framework for the hierarchic organisation and classification of document collections. *Journal of Intelligent Information Systems*, 18(2):153–172, 2002. Special Issue on Automated Text Categorization.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

H.-C. Wang, J. Badger, P. Kearney, and M. Li. Analysis of codon usage patterns of bacterial genomes using the self-organizing map. *Molecular Biology and Evolution*, 18: 792–800, 2001.

F. Wang-Johanning, A. R. Frost, B. Jian, R. Azerou, D. W. Lu, D.-T. Chen, and G. L. Johanning. Detecting the expression of human endogenous retrovirus E envelope transcripts in human prostate adenocarcinoma. *Cancer*, 98(1):187–197, 2003.

J. Watson and F. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171(4356): 737–738, 1953.

E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side information. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*. MIT Press, 2003.

Z. R. Yang and K.-C. Chou. Mining biological data using self-organizing map. *Journal of Chemical Information and Modeling*, 43(6):1748–1753, 2003.

K. Yeung, C. Fraley, A. Murua, A. Raftery, and W. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17:977–987, 2001.

K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17:763–774, 2001.

J.-M. Yi, T.-H. Kim, J.-W. Huh, K. S. Park, S. B. Jang, H.-M. Kim, and H.-S. Kim. Human endogenous retroviral elements belonging to the HERV-S family from human tissues, cancer cells, and primates: expression, strucure, phylogeny and evolution. *Gene*, 342:283–292, 2004.