Methodology for exploring and communicating molecular characteristics of disease

Dennis Alexander Kostka

October 2006

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Gutachter:
Prof. Dr. Martin Vingron
Prof. Dr. Peter Martus

1. Referent: Prof. Dr. Martin Vingron
2. Referent: Prof. Dr. Peter Martus
Tag der Promotion: 11ter Dezember 2006

Contents

Acknowledgements

While working on this thesis I was part of the *Computational Diagnostics Group* within the Department of Computational Molecular Biology at the Max Planck Institute for Molecular Genetics, Berlin. I wish to thank all the people who contributed to the exceptional environment there for the good time I had. I only want to single out my two office-mates, *Stefanie Scheid* and *Florian Markowetz*. Thanks again, guys.

My supervisor *Rainer Spang* provided the topic as well as discussions, insights, a considerable amount of slack and, most important, ongoing support: Thanks a lot. *Martin Vingron* and *Knut Reinert*, who both participated in my PhD committee, provided helpful comments and suggestions. I am grateful for their time and guidance. I am also thankful that *Abha Singh Bais*, *Stefan Bentink*, *Jochen Jäger*, *Claudio Lottaz*, *Florian Markowetz* and *Stefan Roepke* read, commented on and discussed about the manuscript.

Finally I would like to mention my parents *Inge* and *Arno Kostka*. It goes without saying that without them I would not have been here. But ongoing encouragement and support does not. This last line, simply, is to *Abha*.

Berlin–October 2006,
Dennis Kostka

Introduction

In my thesis I present two methodological contributions to the field of microarray data analysis. Microarrays are miniature devices built to simultaneously measure the abundance of messenger ribonucleic acid (mRNA) for large numbers of genes. Such data characterizes cells on a molecular level, offering new possibilities as well as challenges in analysis and interpretation. In the following I motivate the use of microarrays in a clinical setting, introduce the underlying technology and present an outline of the thesis.

Microarrays characterize disease

All cells in the human body contain the same genetic information, stored in the form of deoxyribonucleic acid (DNA). The DNA molecule consists of two long polymer chains, two strands, that form a double-helix structure. Each strand is a sequence of nucleotides containing one of the bases adenine (A), guanine (G), cytosine (C) or thymine (T). The two strands are exactly *complementary*; that is, the sequence of one strand is completely determined by the other: A always pairs with T and G with C (see left picture in Figure I.1).

Cells acquire different characteristics by utilizing different functional units of the DNA (different *genes*). Distinct parts of the DNA are made accessible, *transcribed* to mRNA and *translated* into protein (see Figure I.1). This leads to the different functionality and "behaviour" of the various types of cells. If a gene is transcribed to mRNA, we also say the gene is *expressed*. Transcription is a necessary prerequisite for a gene, or rather its associated protein, to contribute to the attributes of a cell. The mRNA molecules are sequences of the same nucleotides making up the DNA, with the exception that thymine (T) is replaced by uracil (U). Consequently, the pairings A–U and C–G are preferred. In this way, the nucleotide sequence of an expressed gene is transferred to the mRNA (two left pictures of Figure I.1) and determines the protein to be produced.

Microarrays measure the abundance of messenger RNA. Remarkably, the expression of *tens of thousands* of genes is measured at a time, providing an almost comprehensive picture of transcriptional activity. In case a disease is caused by abnormal "behaviour" of cells, this is bound to be reflected on a transcriptional level. Microarrays can then be used to find and classify such reflections. This kind of information can be used to elucidate the biology of disease mechanisms and to improve diagnosis and prognosis of disease.

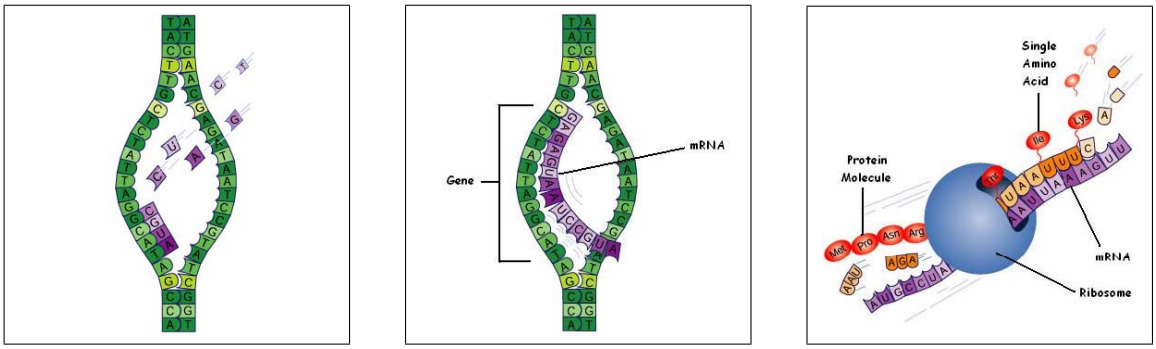

Figure I.1: The central dogma of molecular biology: *DNA makes RNA makes protein. In the left picture DNA (green) is made accessible for transcription (the double strand is broken apart). In the middle picture a gene has been transcribed from DNA to RNA (purple); we also say the gene is* expressed. *It is then transported to the ribosome and translated into protein (right picture). Pictures have been reproduced from [1].*

Microarray technology

While microarrays can be realized using different technical platforms, the general concept remains similar. Specifically designed *probes*, short DNA strands, are placed on an inert substrate. This makes up the microarray device. To measure gene expression, RNA is extracted from the tissue of interest, amplified and fluorescence-labeled. Then the RNA is washed over the array, where it hybridizes to *complementary* DNA probes (see also Figure I.2). As the composition of the probes is known, so is the location on the array where corresponding complementary RNA fragments will bind. Excess RNA is washed off and fluorescence intensity is measured with a laser scanner. The signal strength is related to the number of labeled RNA fragments present and resembles gene expression.

In the following we take a closer look at Affymetrix® GeneChip® technology, because data we analyze have been generated with this kind of microarray. GeneChips are *oligonucleotide* microarrays. The probes on the substrate are short DNA molecules (oligomers), each consisting of 25 nucleotides. They are synthesized directly on the substrate using a photolithographic process. The substrate is a 1.25 cm by 1.25 cm square, subdivided into many smaller squares, the *probe cells* or *features*. The side-length of a feature can be as small as five micron (5/1000 millimeter); each feature contains millions of copies of a unique type of probe. The probes are chosen to be specific for a gene, that is only one gene contains a nucleotide-sequence of length 25 perfectly matching the probe. Additionally, several probes are combined to optimally represent an entire gene. Usually about eleven probes are assigned to the same gene, and the corresponding features are called a *probe set*. Designing good probes is non-trivial and known as the *probe selection problem* [80, 109].

To measure gene expression RNA is extracted from a tissue sample. In a first processing step the RNA is reverse transcribed to complementary DNA. This step is necessary, as the following in-vitro-transcription produces *biotin labeled* RNA matching the original tem-

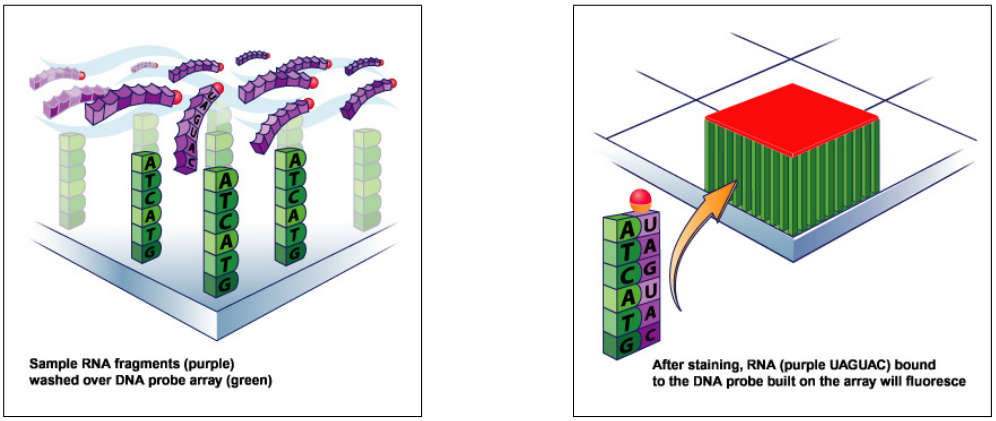

Figure I.2: The principle of microarray technology. *Biotin labeled RNA fragments (purple) are washed over the microarray (left picture). They preferentially hybridize to* complementary *DNA probes (green, lower left corner of the right picture). After treatment with fluorescent stain that sticks to the biotin molecule, RNA fragments binding to probes can be recognized by a laser scanner. Intensity of the signal is related to the number of RNA fragments and resembles gene expression. Pictures have been reproduced from [1].*

plate. The biotin molecules later serve as docking points for fluorescent molecules and enable the quantification of RNA on the array. Then the labelled RNA is fragmented to smaller pieces and washed over the chip to hybridize. This is shown in the left picture of Figure I.2. The labelled RNA fragments (purple) preferentially hybridize to complementary probes on the respective feature. The number of fragments hybridizing to a feature resembles the abundance of this type of RNA. After excess fragments are washed off, a fluorescent molecule that binds to the biotin is used to mark the hybridized RNA. A confocal laser scanner is used to read out the information (right picture of Figure I.2); features with many bound RNA fragments produce high signal intensities while features with few fragments give a weak signal . For each feature, the corresponding intensity resembles the amount of associated RNA present in the tissue sample. Intensity values are digitized and form the basis of subsequent analysis. Since each feature produces its own signal and resembles a unique probe, this data is also called *probe level data*. To reliably quantify the expression of a gene, information from the different features of each probeset is aggregated. This process is called *probeset summary* and it is part of a *preprocessing* step generally applied to the raw probe level data; more details can be found in Chapter 2. The aggregated data is also called *gene level data*.

Thesis outline

Microarray data data characterizes cells on the transcriptional level. Prominent applications of microarray technology in a clinical setting are the molecular diagnosis of patients and the discovery of disease subtypes by patient stratification (clustering). Lists of differentially expressed genes are often used to guide biological intuition. In general,

the data can be utilized to infer novel biological hypotheses by means of pattern mining and to refine or confirm existing knowledge. This thesis contains methodological contributions to both settings. It is composed of three chapters.

The first chapter describes statistical learning techniques, which are frequently applied to microarray data with the goal of obtaining rules for molecular diagnosis. The focus lies on characteristics arising from the specific nature of high dimensional microarray data. This chapter *concisely integrates concepts, algorithms and practical aspects* of microarray data analysis that are usually found in distinct fields of the literature. It provides the theoretical foundation of the other chapters.

The second chapter is concerned with the unambiguous documentation of a diagnostic molecular signature or, equivalently, with the unequivocal characterization of disease or subtype of disease. The motivation to address documentation and communication of molecular signatures is a practical one: Microarray based gene expression signatures have the potential to be powerful tools for patient stratification and diagnosis of disease. But before they can affect clinical practice they need to be communicated to other health care centers with data for independent validation [130].

External validation of a signature can only be meaningful if the new data is transformed to a scale compatible with the original one the signature is tuned to. This scale, in turn, depends on the initial preprocessing applied in the signature deriving study. It needs to be communicated alongside with the signature. Chapter two formalizes this requirement and contains scale adjusting transformations for two popular preprocessing schemes, `rma` [72] and `vsn` [67, 68]. In both cases data dependent parameters that determine the scale are identified and algorithms to adjust external data are provided. Using eight clinical microarray data sets I am able to show significantly increased consistency and stability of molecular diagnoses as compared to standard documentation procedures. This underlines the key point of the chapter: *Data preprocessing has to be taken into account* when documenting molecular characteristics of disease. In case of `vsn` the scale adjusting algorithm comprises a maximum likelihood estimation of tansformation parameters; the usefulness of the per-chip likelihood score as an indicator for the compatibility of external data to the core study is also assessed.

The third chapter introduces the `dcoex` algorithm, a method designed to utilize microarray data to reveal groups of genes losing coregulation between two phenotypes. Information about differentially coregulated genes can not only provide a molecular characterization of the phenotypes; it also provides focused information which is useful to generate hypotheses about biological mechanisms underlying the phenotypical differentiation. This chapter introduces the concept, implements an algorithm for detection and demonstrates the biological plausibility of *differentially coexpressed genes*.

Since coregulation cannot be measured on microarray data, the objective is to find groups of genes coexpressed in one type of samples that lose their coexpression for a second phenotype. That is, genes with pronounced differences in their dependency structure (conditional on the phenotype) are sought. The `dcoex` algorithm identifies

such groups of genes by minimizing a differential coexpression score S. The objective is stated as a binary polynomial fractional program [23], which we approach heuristically by a stochastic descent algorithm. By exploiting S in deriving efficiently computable criteria for score reduction we are able to quickly identify downhill steps. We demonstrate the the improvement over a naive descent strategy and apply the algorithm to simulated and real data. In a data set on childhood leukemia [157] we find a biologically plausible group of genes differentially coexpressed between cytogenetically normal children and children bearing a Philadelphia chromosome. After assessing robustness and statistical significance of our findings we conclude that dcoex constitutes a new analysis tool enabling the exploration of differential coexpression patterns.