Technical University of Crete
School of Electronic and Computer Engineering

Exploration of disease-specific biomarkers in Cancer Research by integrating biological knowledge and high throughput data

by

Stelios Sfakianakis

A dissertation submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Chania, July 2016

The Dissertation of Stelios Sfakianakis
is approved:

Professor Michalis Zervakis, Chair
School of Electrical and Computer Engineering,
Technical University of Crete

Professor Emeritus Stavros Christodoulakis
School of Electrical and Computer Engineering,
Technical University of Crete

Associate Professor Georgios Karystinos
School of Electrical and Computer Engineering,
Technical University of Crete

Professor Minos Garofalakis
School of Electrical and Computer Engineering,
Technical University of Crete

Assistant Professor Daphne Manousaki
School of Electrical and Computer Engineering,
Technical University of Crete

Dr. Dimitris Kafetzopoulos
Institute of Molecular Biology and Biotechnology,
Foundation for Research & Technology - Hellas (FORTH)

Professor Manolis Tsiknakis
Department of Informatics Engineering,
Technological Educational Institute of Crete

Abstract

Motivation In the post-genome era high throughput technologies like DNA arrays provide the massive expression profiling of thousands of genes and have become a powerful tool for the state of the art scientific research especially in relation to the treatment of complex, multifactorial diseases such as cancer. Breast cancer is widely known as the most common malignancy in women worldwide and the presents the second highest mortality rate. In breast cancer patients, it is not the primary tumour, but its metastases at distant sites that are the main cause of death. To establish a metastasis, tumour cells have to invade their surrounding host tissue, enter the circulatory blood stream, arrest in capillary beds of distant organs, invade the host tissue and proliferate. These so-called circulating tumor cells (CTCs) are now considered to play a key role in metastasis and their study and characterization is becoming pivotal for the detection of cancer and metastasis at an early stage.

Goal During the recent couple of years a number of techniques have been developed for the isolation of CTCs but research efforts encounter many challenges such as the scarcity of CTCs in the patients' blood. The CTC detection methods are usually based on the physical properties of CTCs (e.g. filtration based on size), or by using specific antibodies able to recognize specific tumor markers such as the epithelial cell adhesion molecule (EpCAM). An alternative, interesting approach is to follow a data-driven methodology that through the use of statistics and computational techniques aims to identify differences and similarities between the blood and tissue samples of cancer patients and healthy populations. Potential discoveries in this endeavor can provide answers for the molecular characterization of metastatic breast cancer and the presence of CTCs.

Approach In order to proceed to a statistically sound genomic classification of tissue and blood of breast cancer patients a data integration approach has been designed. A large compendium of publicly available gene expression data sets has been brought together and carefully merged in order to overcome

study specific biases or platform related technical variations. This integration methodology is then followed by a number of statistical comparisons between the different in origin (blood or tissue) or in disease status (cancerous or healthy) samples in order to reveal potential "biomarkers" for each case. These biomarkers are genes that exhibit different behavior (e.g. over-expression) in the aforementioned comparisons but in order to increase the sensitivity the sets of discriminating genes are intersected and a common subset is identified. The unique set of genes derived is then related to well curated sources of biological knowledge, such as biological networks, and subjected to novel algorithmic procedures so as to establish the underlying biological foundation and to further elicit features (genes) for the supervised and unsupervised classification of breast cancer patients.

Contributions The key deliverables of this work is the identification of a 27-genes signature as potential markers for the characterization of CTCs and the metastatic cascade, and a number of computational methods along their findings that take advantage of existing biological knowledge to fine tune the derived signature for the supervised or unsupervised classification of patient samples, as follows:

- Following the methodology that was briefly described above, 9 different data sets publicly available from the Gene Expression Omnibus (GEO) database were collated, assembled, and integrated, yielding more than 800 samples of gene expression values. The subsequent statistical analysis and integration of results produced a "genes signature" of 27 genes as candidate biomarkers related to the presence of CTCs and two of them (CXCR4 and JUNB) were in fact found to be really CTC-related in a biological "bench-top" experiment, effectively confirming the statistical findings.

- The next question is whether the genes participating in the derived signature are related and how they affect each other. By introducting biological networks, these questions are translated to the Steiner tree problem in graphs. This formulation and the corresponding solution in a high quality protein-protein interaction network reveal the shortest interconnect for the genes in our signature and enhance it with additional central genes along the interconnecting paths.

- The ability of the 27-genes signature as a guiding, feature extraction and generation tool for the classification or "clustering" of patient samples is then examined, again using the underlying network information. We first introduce a two-level classification scheme that uses as base classifiers the "neighborhoods" of the 27 genes, which were induced using random walks in the biological graph. Secondly, for the problem of classification into unknown categories (“clustering”) or the identification of new groups of patients, a model-based statistical algorithm is adapted to use the “neighbors” of the 27 genes that effectively alleviates the problem of dimensionality.

It always seems impossible until it's done.

Nelson Mandela

- Πώς πρέπει ν' αγαπούμε τους ανθρώπους;
- Μοχτώντας να τους φέρουμε στον σωστό δρόμο
- Και ποιος είναι ο σωστός δρόμος;
- Ο ανήφορος

Νικος Καζαντζακης

Στη Βούλα, το Γιώργο, και το Γιάννη

Acknowledgments – Ευχαριστίες

I want to thank first of all my supervisor Prof. Michalis Zervakis for giving me this opportunity to learn so much about research and introducing me to this important field. His guidance and experience were the most valuable tools for completing this work.

I am deeply indebted to Dr. Katerina Bei for her encouragement and invaluable help throughout this project. I couldn't have come this far without her patience, knowledge, insights, and commitment.

I want also to express my sincere gratitude to my advisory and final examination committee for their comments and support, especially Dr. Manolis Tsiknakis who was the person that initially urged me in pursuing a PhD, and Dr. Dimitris Kafetzopoulos who gave important advice on the formulation of the research question.

Δεν ϑα μπορέσω ποτέ να ευχαριστήσω αρκετά τους γονείς μου που μου έμαϑαν να εργάζομαι με φιλότιμο ϰαι που πάντα ήταν δίπλα μου να με στηρίξουν.. Τις αδελφές μου Βαρβάρα, Στέλλα, ϰαι Ροδούλα για την υποστήριξη ϰαι την αγάπη τους.. Τους γονείς της συζύγου μου, Ελένη ϰαι Γιάννη Καρπαϑιωτάκη, για τη βοήϑεια ϰαι το ενδιαφέρον τους. Τέλος, εκφράζω την αμέριστη ευγνωμοσύνη μου στην αγαπημένη μου σύζυγο Βούλα για τη συμπαράσταση ϰαι την υπομονή της όλα αυτά τα χρόνια ϰαι στους αγαπημένους μας γιους, το Γιώργο ϰαι το Γιάννη. Αυτή η διατριβή είναι αφιερωμένη σε αυτούς.

Contents

Chapter 1

Introduction

Contents

Cancer is a highly complex and heterogeneous disease which involves a succession of genetic changes that eventually results in the conversion of normal cells into cancerous ones. It is obvious that a complete understanding and knowledge of these processes requires the integration and analysis of massive amounts of data as is being collected from current genomic, proteomic and metabolomic platforms [Ge et al., 2003]. But it is not only the multiplicity of the factors (and cellular levels) contributing to a particular disease framework that imposes approaching the problem in a systematic way. Even for Mendelian genetic disorders, nearly all of which have now been correlated with a specific gene or set of genes [Hoh and Ott, 2004] due to remarkable advances in gene mapping and bioinformatics, the relationship between genotype and phenotype is not as simple as expected (and/or currently treated) [Scriver and Waters, 1999]. Because our knowledge of this domain is still largely rudimentary, investigations are now routinely moving from being "hypothesis driven" to being "data-driven" with analysis based on a search for biologically relevant patterns. These technological advances have created enormous opportunities for accelerating the pace of science. In this context, exploratory analyses is the process of generating hypotheses that are later supported (or not) by the data (e.g. hypothesis: gene x is responsible for a side effect of drug y) [Sfakianakis et al., 2010a].

This data driven approach is the main theme of the current thesis. We are interested in the hypothesis generation through the computational analysis of high throughput, gene expression data, focusing, mainly, on the molecular characterization of Breast Cancer and its metastatic cascade. Breast cancer is widely known as the most common malignancy within the population of western Caucasian women and is also the second most common type of cancer with fatal outcome in female populations [Siegel et al., 2015]. In breast cancer patients, it is not the primary tumor, but its metastases at distant sites that are the main cause of death. A prominent theory on the cancer progression and metastasis is the invasion of tumour cells into surrounding stroma in order to subsequently intravasate and enter the blood circulation, effectively, becoming *Circulating Tumor Cells* (CTCs) , which may represent "metastatic intermediates" [Valastyan and Weinberg, 2011]. The importance of CTCs, therefore, for early detection of cancer and metastasis has been prominently reported in the last couple of years and a multitude of approaches and techniques for their detection and characterization have been proposed by the scientific community. Instead, what we followed is a multi-comparison statistical methodology that aims to identify differences and similarities between the blood and tissue samples of cancer patients and healthy populations. Potential discoveries in this endeavor can provide answers for the molecular characterization of metastatic breast cancer and the presence of CTCs.

Towards this aim, high throughput technologies like DNA microarrays provide the massive expression profiling of thousands of genes and have become a powerful tool for the state of the art scientific research [Van't Veer et al., 2002].

Nevertheless, whereas in traditional applications of pattern recognition and data mining there are large number of samples with a small feature space, in the area of bioinformatics the mass of data produced exhibit the exact reverse characteristics: small sample size with a large number of features, a problem commonly referred as the "curse of dimensionality" [Clarke et al., 2008]. In this environment standard statistical and machine learning methods are likely to over-fit the structures in the data, and the presence of "noise" puts statistical analysis and inference under fire. To alleviate this problem, in addition to standard computational techniques that are successfully employed in other domains with similar characteristics (e.g. image recognition, computer vision, etc), the use of prior biological knowledge can be invaluable as a dimensional reduction technique, a heuristics generating strategy, as well as a means for the validation of results or for the generation of further hypotheses.

Therefore, the objective of this chapter is primarily to lay the ground and present the setting of our research. We briefly provide some background information both from the biological domain point of view and from the computational and technical viewpoints. Subsequently, we present the main roadmap of our work, the objectives and the main contributions. The succeeding chapters expand on these topics.

1.1 Background on Biology and Breast Cancer

Since the focus of our work is within the confines of bioinformatics [Baldi and Brunak, 2001], we start by providing some background information on the domain: the molecular biology with emphasis on molecular genetics, the pathology of Breast Cancer, and the circulating tumor cells as the mechanisms for the creation of secondary tumors and metastases in cancer patients. Due to the space limitations and the urge to proceed to the main subject of the thesis, the exposition is brief and, of necessity, incomplete. The referenced publications can provide a more thorough account of this material to the interested readers.

1.1.1 DNA, RNA, and the "central dogma" of Biology

Cells are the fundamental working units of every living system. Inside their nuclei (in the case of eucaryotes), a molecular sequence known as *deoxyribonucleic acid*, or DNA, conveys the instructions needed to direct the activities of the cells [Brown, 2006]. In fact, DNA is the primary molecule of inheritance as it carries most of the genetic instructions used in the growth, development, and functioning of nearly all organisms. It was discovered in 1869 by Swiss physician and biochemist Friedrich Miescher inside the nuclei of human white blood cells (leucocytes) [Dahm, 2008]. In 1953, American biologist James Watson and English physicist Francis Crick concluded that the DNA molecule exists in the form of a three-dimensional double helix, as depicted in Figure 1.1 [Watson et al., 1953]. Another highly important molecule is the *ribonucleic acid* (RNA). This is a single-stranded molecule and its main functionality is the interpretation of the genetic information stored in the DNA.

DNA and RNA share a similar structure. They are composed of a series of nucleotides and each nucleotide has three components: a phosphate group; a pentose sugar (deoxyribose in the case of DNA, ribose in the case of RNA); and a nitrogen-containing base. There are two basic categories of nitrogenous bases: the purines (adenine [A] and guanine [G]), and the pyrimidines (cytosine [C], thymine [T], and uracil [U]). RNA contains only A, G, C, and U (no T), whereas DNA contains only A, G, C, and T (no U). These bases are paired together by forming hydrogen bonds according to the pairing (binding) rules: A always pairs with T in DNA or with U in RNA; G always pairs with C, in the case of both, DNA and RNA. The double stranded form of DNA is kept together through these bindings (Figure 1.1).

The *human genome*, i.e. the total composition of genetic material within a cell, is packaged into larger units known as chromosomes. The chromosomes are physically separate molecules that range in length from about 50 million to 250 million base pairs (Figure 1.2). Human cells contain two sets of chromosomes, one set inherited from each parent. Each cell, except sperm and eggs, contains 23 pairs of chromosomes — 22 "autosomes"[1] (numbered 1 through 22) and one pair of sex chromosomes (XX or XY). Sperm and eggs contain half as much genetic material (e.g., only one copy of each chromosome) [Alliance, 2010].

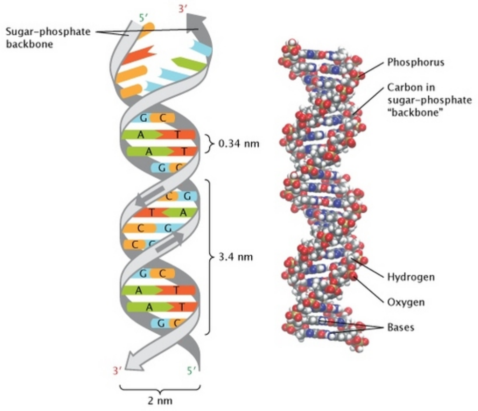


Figure 1.1: The double-strand helix of DNA. Image courtesy of Nature Education.

The generic information encoded by DNA in a cell of an organism guides the functioning of the cell through the creation of *proteins*. Proteins form enzymes and macromolecules active in cellular structure and biochemical processes. The complete set of proteins expressed by an organism at a particular time is

[1] An autosome is a chromosome that is not an "allosome". An allosome is a sex chromosome.

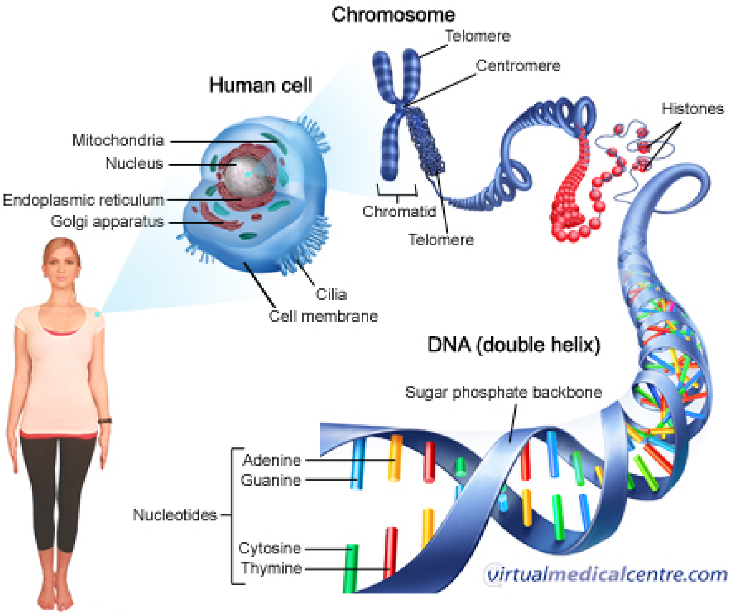


Figure 1.2: DNA is a double helix formed by base pairs attached to a sugar-phosphate backbone. DNA is found inside a special area of the cell called the nucleus in a packaged form that is called a chromosome. DNA is made of chemical building blocks called nucleotides. There are four types of "nucleobases" forming pairs in the DNA helix: adenine (A), thymine (T), guanine (G), and cytosine (C). The sequence of these bases determines what biological instructions are encoded in a strand of DNA. The complete DNA of an individual, her or his genome, contains about 3 billion bases and about 20,000 genes on 23 pairs of chromosomes. (Image courtesy of www.myvmc.com)

called the *proteome* and, in contrast with the genome which is normally stable, it is dynamically influenced by various factors, including internal and external conditions of the cell.

The so called *central dogma of molecular biology* [Crick et al., 1970] defines the production of proteins from DNA through the use of RNA, as illustrated in Figure 1.3. Briefly, a specific sequence of DNA (an eukaryotic protein-coding gene) is transcribed into pre-mRNA by the means of RNA polymerase. This RNA is then usually modified (splicing) by an RNA- protein complex called the spliceosome[2]. Once the pre-mRNA is processed (maturation), the resulting mRNA message is then translated by the ribosome in order to produce proteins (translation). The expression of a particular gene is defined as the level (density) of mRNA produced by the transcription of this gene. The initial product of genome expression is called *transcriptome*, a collection of RNA

[2] Sometimes a pre-mRNA message may be spliced in several different ways, allowing a single gene to encode multiple proteins (alternative splicing).

molecules derived from those protein-coding genes whose biological information is required by the cell at a particular time [Brown, 2006].

A *gene* is a specific segment of a DNA molecule that contains all the coding information necessary to instruct a cell to synthesize a specific product, such as an RNA molecule or a protein. Contained within the gene are segments that we acknowledge as active in the coding process (exons), as well as segments that are non-coding (introns). Each gene also represents a basic unit of a person's biological inheritance from his or her two parents. Genes can be "mapped" because each occupies a specific location (or locus) on a chromosome (out of the 23 pairs of chromosomes), and each chromosome can be specifically identified as well. Currently, the estimated number of human genes is around 20,000 [Ezkurdia et al., 2014].

By definition, human genes function to promote and regulate biological activity that is considered necessary and productive for the functioning of the organism. It is not correct to state that a gene codes for a disease or predisposes a person to a specific disorder. Rather, it is a deleterious mutation in a gene that may predispose a person to a specific disease or disorder [Hernandez et al., 2006].

1.1.2 Breast Cancer

Cancer is the name given to a collection of related diseases (over 100), when some of the body's cells begin to divide without stopping and spread into surrounding tissues. Cancer occurs as a result of mutations, or abnormal changes, in the *genes* responsible for regulating the growth of cells and keeping them healthy. Normally, the cells in our bodies replace themselves through an orderly process of cell growth: healthy new cells take over as old ones die out. But over time, mutations can alter the behaviour of certain genes which results in uncontrolled cell division and the formation of tumors. If untreated, the malignant cells eventually can spread beyond the original tumor to other parts of the body and cause metastasis, which is estimated to be the cause of death to around 90% of the cases [Weigelt et al., 2005].

The term "breast cancer" refers to a malignant tumor that has developed from cells in the breast. Usually breast cancer either begins in the cells of the lobules, which are the milk-producing glands, or the ducts, the passages that drain milk from the lobules to the nipple. Less commonly, breast cancer can begin in the stromal tissues, which include the fatty and fibrous connective tissues of the breast. Breast cancer is the most common form of cancer in women, rarely affecting also men [Anderson et al., 2010], with an estimated 234,190 new cases and 40,730 deaths in 2015 in the United States alone [Siegel et al., 2015] (Figure 1.4). It was first documented in ancient Egyptian writings,

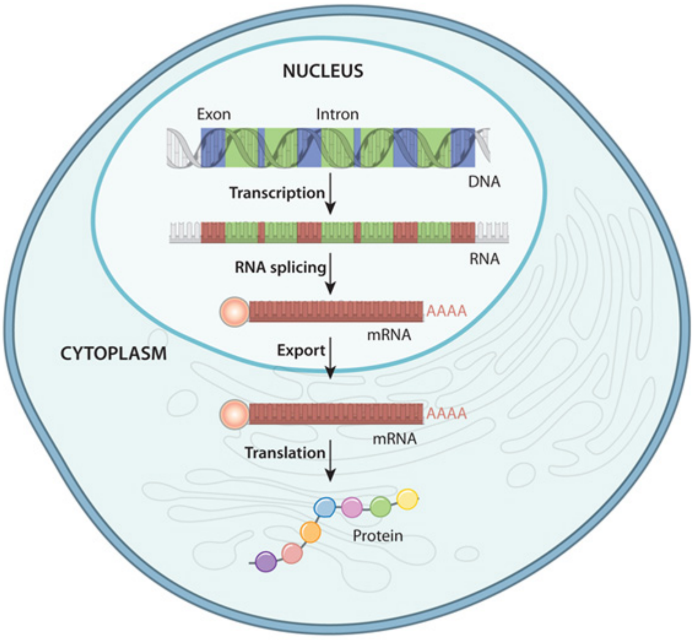
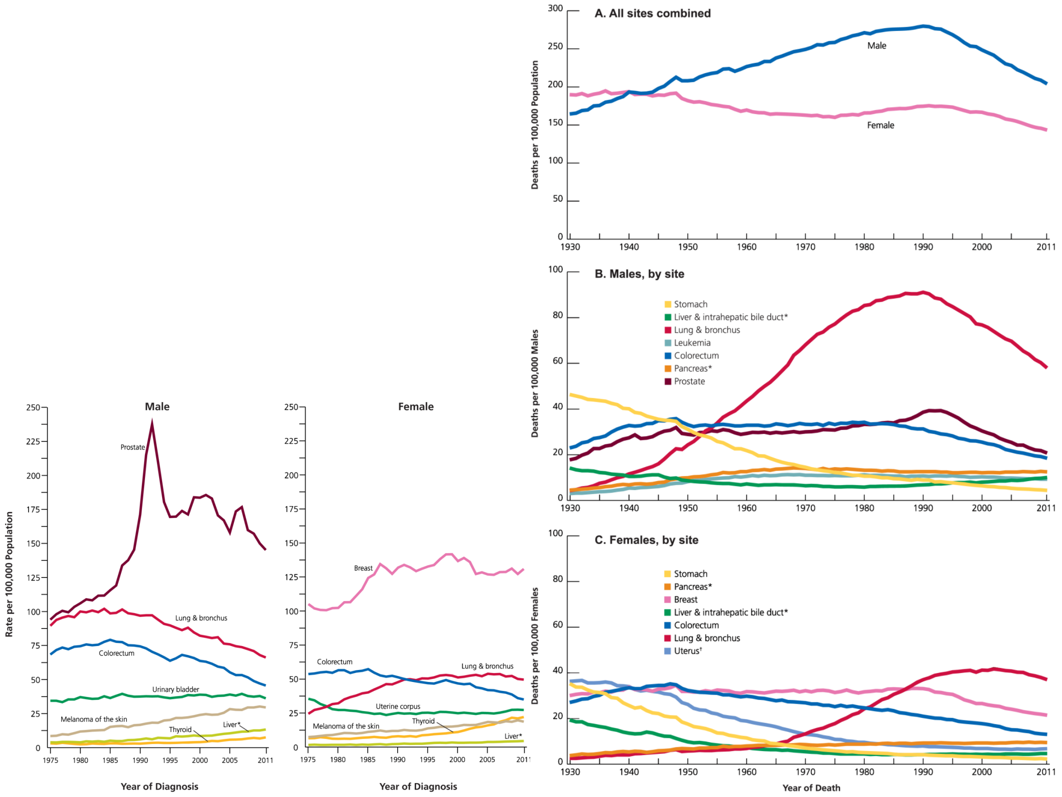


Figure 1.3: The flow of information from DNA to proteins in a eucaryote organism. The coding and noncoding regions of DNA are transcribed into messenger RNA (mRNA) by the enzyme RNA polymerase. The introns are removed during the initial mRNA processing and the remaining exons are then spliced together. The spliced mRNA molecule (red) exports out of the nucleus through and when it arrives in the cytoplasm, it is translated to a protein. Image courtesy of Nature Education.

the Edwin Smith Papyrus (copy of trauma surgery) where 8 cases of breast malignancy were recorded [Lakhtakia, 2014].

Breast cancer's treatment options include local therapy, such as surgery and radiation, and systemic therapy e.g. chemotherapy. The choice of the specific treatment plan depends on a number of factors, such as the age, the size and location of the tumor, and, of course, the stage of the cancer. Stage I and stage II are early stages of breast cancer and are usually treated by mastectomy and/or radiotherapy. At stage III the tumor is large and the cancer has spread to lymph nodes or other tissues near the breast, while at stage IV the cancer has metastasized beyond the breast and underarm lymph nodes to other organs and parts of the body. The most frequent metastatic sites are the bones, the lungs, the skin, and the brain.

Although treatments like hormone therapy and chemotherapy can slow down the spread of cancer or relieve the symptoms, metastatic breast cancer, unfortunately, remains incurable and therefore early detection is very important. A screening mammogram, which is a type of x-ray, is the best tool available for



(a) Incidence Rates *(b) Death Rates*

Figure 1.4: Trends in cancer rates for selected sites by sex, United States, 1930 to 2011. (Reprinted with permission from [Siegel et al., 2015].)

finding breast cancer early, before symptoms appear. Thanks to the routine use of those screening mammograms in developed countries, more and more women diagnosed with breast cancer are detected at an early stage. Despite early detection and initial treatment, approximately 40% of the diseased women will develop distant metastasis, i.e. development of new tumors in different organs [Weigelt et al., 2005]. Most recurrences appear within the first 2 or 3 years after treatment, but breast cancer can recur 10 years or more after the initial diagnosis [Harris et al., 2012].

Due to the prevalence of breast cancer and its importance for public health, there is a vast interest on its understanding and treatment in recent years. Although currently scientists do not have yet enough information to provide definite answers, a number of risk factors like the age (women over the age of 60 are at a higher risk), ethnicity (more frequent in Caucasian women), and family history have been reported. Finally, research has identified the relation of certain genetic alterations (mutations) in genes such as BRCA1 and BRCA2 to the emergence of breast cancer. Based on these findings, the importance of genetic testing becomes more and more acknowledged [Jolie, 2013].

1.1.3 Circulating Tumor Cells

Metastatic disease is responsible for over 90% of cancer deaths [Fidler, 2002, Wittekind and Neid, 2005]. During the first stages of the metastatic cascade, cancer cells escape from a primary tumor mass and intravasate allowing their lymphohematogenous dissemination to distant sites of the body. Most of these "circulating tumor cells" (CTCs) that depart the primary tumor will die, whereas as few as 0.01% of CTCs are likely to give rise to metastases, as suggested by pre-clinical models [Chaffer and Weinberg, 2011]. Once cancer cells extravasate in anatomically distant organs, they can be found as single cells or small number of clustered cells referred to as "disseminated tumor cells" (DTCs) [Zhe et al., 2011, Massagué and Obenauf, 2016] (see also Figure 1.6.)

Early findings

The hypothesis that circulating tumour cells (CTCs) are a fundamental prerequisite to metastasis was first proposed in the mid 19th Century by Thomas Ashworth, an Australian pathologist [Ashworth, 1869], when he wrote (see Figure 1.5 for the full article):

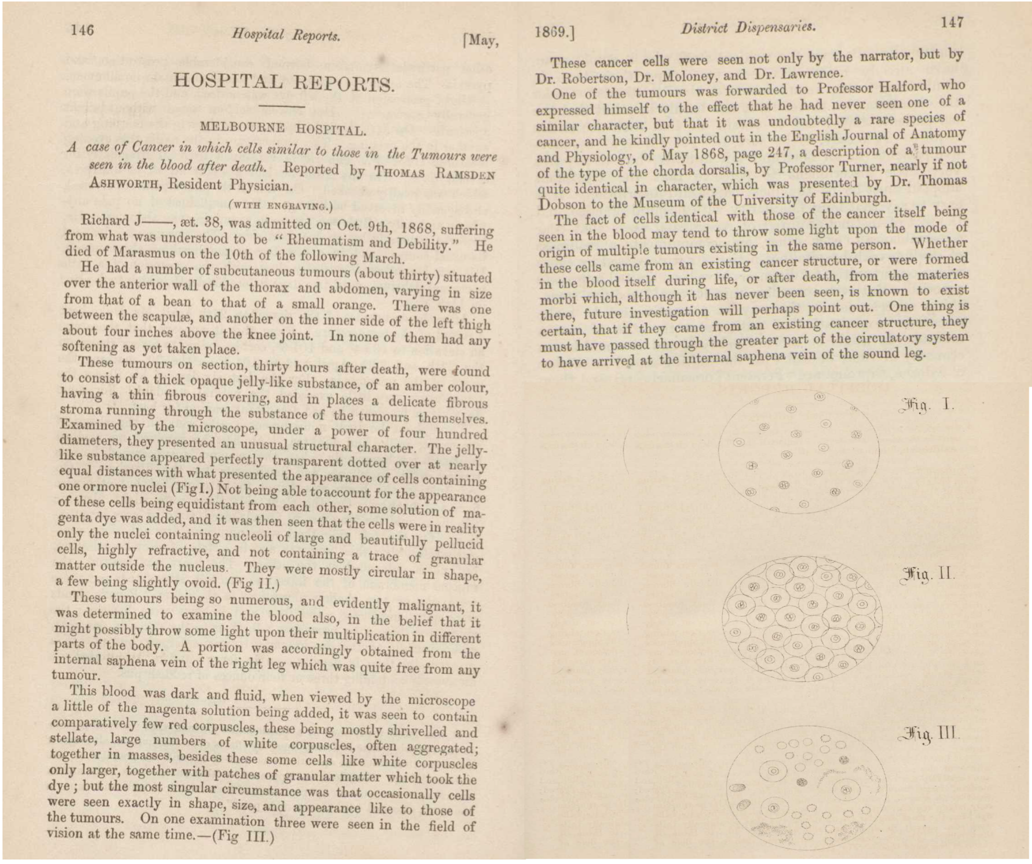
> *The fact of cells identical with those of the cancer itself being seen in the blood may tend to throw some light upon the mode of origin of multiple tumours existing in the same person.*

Progression-free and overall survival

Progression-free survival (PFS) is defined as the time elapsed between treatment initiation and tumor progression or death from any cause, with censoring of patients who are lost to follow-up [Green et al., 2012]. On the other hand, Overall survival (OS) is based on death from any cause, not just the condition being treated.

Increasing evidence suggests that circulating tumor cells (CTCs) in the peripheral blood are associated with reduced progression-free survival (PFS) and overall survival (OS) in metastatic disease [Bidard et al., 2008, Hayes et al., 2006, Botteri et al., 2010, Cristofanilli et al., 2004]. Whereas the detection of CTCs before the start of a new treatment has been associated with poor prognosis, the enumeration of CTCs shortly after the initiation of therapy provides additional information regarding treatment response [Cristofanilli et al., 2004, Hayes et al., 2006].

Cancer patients have only between 5 and 50 CTCs per teaspoon of blood (or approximately 1 CTC per billion blood cells!!), so their presence is dwarfed



146 *Hospital Reports.* [May,

HOSPITAL REPORTS.

MELBOURNE HOSPITAL.

A case of Cancer in which cells similar to those in the Tumours were seen in the blood after death. Reported by Thomas Ramsden Ashworth, Resident Physician.

(with engraving.)

Richard J——, æt. 38, was admitted on Oct. 9th, 1868, suffering from what was understood to be "Rheumatism and Debility." He died of Marasmus on the 10th of the following March.

He had a number of subcutaneous tumours (about thirty) situated over the anterior wall of the thorax and abdomen, varying in size from that of a bean to that of a small orange. There was one between the scapulæ, and another on the inner side of the left thigh about four inches above the knee joint. In none of them had any softening as yet taken place.

These tumours on section, thirty hours after death, were found to consist of a thick opaque jelly-like substance, of an amber colour, having a thin fibrous covering, and in places a delicate fibrous stroma running through the substance of the tumours themselves. Examined by the microscope, under a power of four hundred diameters, they presented an unusual structural character. The jelly-like substance appeared perfectly transparent dotted over at nearly equal distances with what presented the appearance of cells containing one ormore nuclei (Fig I.) Not being able toaccount for the appearance of these cells being equidistant from each other, some solution of magenta dye was added, and it was then seen that the cells were in reality only the nuclei containing nucleoli of large and beautifully pellucid cells, highly refractive, and not containing a trace of granular matter outside the nucleus. They were mostly circular in shape, a few being slightly ovoid. (Fig II.)

These tumours being so numerous, and evidently malignant, it was determined to examine the blood also, in the belief that it might possibly throw some light upon their multiplication in different parts of the body. A portion was accordingly obtained from the internal saphena vein of the right leg which was quite free from any tumour.

This blood was dark and fluid, when viewed by the microscope a little of the magenta solution being added, it was seen to contain comparatively few red corpuscles, these being mostly shrivelled and stellate, large numbers of white corpuscles, often aggregated; together in masses, besides these some cells like white corpuscles only larger, together with patches of granular matter which took the dye; but the most singular circumstance was that occasionally cells were seen exactly in shape, size, and appearance like to those of the tumours. On one examination three were seen in the field of vision at the same time.—(Fig III.)

1869.] *District Dispensaries.* 147

These cancer cells were seen not only by the narrator, but by Dr. Robertson, Dr. Moloney, and Dr. Lawrence.

One of the tumours was forwarded to Professor Halford, who expressed himself to the effect that he had never seen one of a similar character, but that it was undoubtedly a rare species of cancer, and he kindly pointed out in the English Journal of Anatomy and Physiology, of May 1868, page 247, a description of a tumour of the type of the chorda dorsalis, by Professor Turner, nearly if not quite identical in character, which was presented by Dr. Thomas Dobson to the Museum of the University of Edinburgh.

The fact of cells identical with those of the cancer itself being seen in the blood may tend to throw some light upon the mode of origin of multiple tumours existing in the same person. Whether these cells came from an existing cancer structure, or were formed in the blood itself during life, or after death, from the materies morbi which, although it has never been seen, is known to exist there, future investigation will perhaps point out. One thing is certain, that if they came from an existing cancer structure, they must have passed through the greater part of the circulatory system to have arrived at the internal saphena vein of the sound leg.

Figure 1.5: The complete article of [Ashworth, 1869] where it was first observed the similarity between the cancer cells and the cells found in the blood of diseased patient. The article can be found at http://hdl.handle.net/11343/23133 (accessed on February 15, 2016.)

by blood cells (Figure 1.7). However, in the past decade emerging technologies have, for the first time, allowed the isolation of CTCs from patients' blood samples. Some methods, among the first established, rely on the cells' physical properties. When a blood sample settles or is spun in a centrifuge, red blood cells, white blood cells, and other components of blood separate into layers. Based on their buoyancy, CTCs can be found in the white blood cell fraction. Then, because CTCs are generally larger than white blood cells, a size-based filter can divide the cell types.

Dissemination of cancer cell in the blood corresponds to one of the first step of the metastatic process (when cancer cells detach from the primary tumor and intravasate), while micrometastasis in a distant organ, like the bone marrow, reflects a more advanced stage [Bidard et al., 2013]. In early breast cancer, CTC count is also a prognostic biomarker, not correlated with the other usual prognostic factors [Bidard et al., 2016]. Considering the prognostic effect of

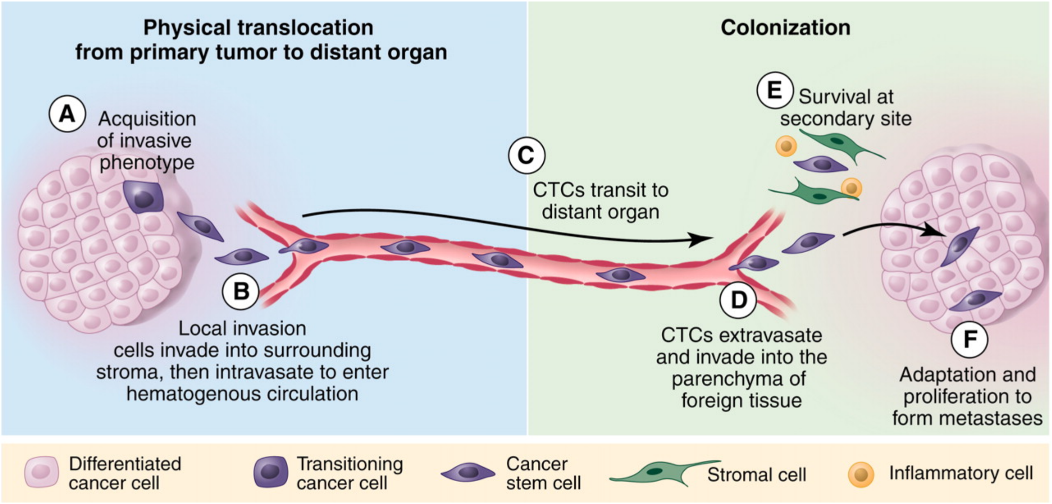

Physical translocation from primary tumor to distant organ
Colonization
A
Acquisition of invasive phenotype
B
Local invasion cells invade into surrounding stroma, then intravasate to enter hematogenous circulation
C
CTCs transit to distant organ
D
CTCs extravasate and invade into the parenchyma of foreign tissue
E
Survival at secondary site
F
Adaptation and proliferation to form metastases
Differentiated cancer cell
Transitioning cancer cell
Cancer stem cell
Stromal cell
Inflammatory cell

Figure 1.6: The metastatic cascade. Metastasis can be envisioned as a process that occurs in two major phases: (i) physical translocation of cancer cells from the primary tumor to a distant organ and (ii) colonization of the translocated cells within that organ. (A) To begin the metastatic cascade, cancer cells within the primary tumor acquire an invasive phenotype. (B) Cancer cells can then invade into the surrounding matrix and toward blood vessels, where they intravasate to enter the circulation, which serves as their primary means of passage to distant organs. (C) Cancer cells traveling through the circulation are CTCs. They display properties of anchorage-independent survival. (D) At the distant organ, CTCs exit the circulation and invade into the microenvironment of the foreign tissue. (E) At that foreign site, cancer cells must be able to evade the innate immune response and also survive as a single cell (or as a small cluster of cells). (F) To develop into an active macrometastatic deposit, the cancer cell must be able to adapt to the microenvironment and initiate proliferation. (Reprinted with permision from [Chaffer and Weinberg, 2011])

CTCs, it is of interest to determine how the effects of such single cells are diluted within the peripheral blood circulation. Clinical evidence suggests that the number of CTCs before treatment is an independent predictor of progression-free survival (PFS) and overall survival (OS) in patients with metastatic breast cancer (MBC), with a prognostic power independent of and either equivalent or superior to tumor burden and disease phenotype [Cristofanilli et al., 2004, Cristofanilli et al., 2007]. Particularly, Cristofanilli et al. [Cristofanilli et al., 2004] reported that MBC patients with $\geqslant$5 CTCs/7.5 ml of whole blood, as compared with the patients with fewer than 5 CTCs/7.5 ml of blood, had a shorter median PFS and shorter OS. This poor prognosis is confirmed in a pooled meta-analysis of individual patient data from 1,944 MBC patients from 20 studies [Bidard et al., 2014]. This threshold seems also to agree with the data shown in Figure 1.7.

Furthermore, based on clinical data, Giuliano et al. [Giuliano et al., 2014] reported a highly different anatomical distribution of metastatic sites in patients with advanced breast cancer according to their baseline CTC counts ($\geqslant$5 versus <5 CTCs/7.5 ml). Thus, a prevalence of bone involvement or soft-tissue/lymph node involvement is higher in patients with $\geqslant$5 CTCs/7.5 ml or with <5 CTCs/7.5 ml, respectively.

Considering the lower concentration of CTCs detected in non-metastatic compared with metastatic breast cancer, this cut-off point has changed to a value of $\geqslant$1 CTC in order to assess the prognostic role of CTCs. The large pooled analysis of individual data from 3,173 patients with non-metastatic (stage I-III) breast cancer from five breast cancer institutions conducted by Janni et al. demonstrates the significant prognostic relevance of CTCs in primary breast cancer patients, independent of the particular frequency of presence [Janni et al., 2016]. Also, [Maltoni et al., 2015] demonstrated that CTC-positivity in pre-surgery early breast-cancer patients often had negative prognostic features, i.e. large tumor dimension, high proliferation, negative receptor status, lymph node positivity and it was highly correlated with vascular invasion. Thus, the utility of CTCs as a predictive and prognostic marker appears to be of great importance in patients with early breast cancer, since it is assumed that their presence at the time of primary diagnosis could predict early disease recurrence and prevail reduced survival based on preliminary prospective clinical trials [Franken et al., 2012, Lucci et al., 2012].

Taking advantage of the recent achievements in detecting and quantifying rare CTCs (as few as 1 CTC per 10^6 - 10^8 leukocytes) in the peripheral blood of cancer patients, the researchers can explore their clinical effectiveness by real-time monitoring of disease progression and treatment responses based on repeated blood sampling [Joosse et al., 2015, McInnes et al., 2015, Janni et al., 2016]. However, beyond the enumeration of CTCs, the delineation of their gene and/or protein expression profiles, as well as the identification of the biological properties (e.g. processes, pathways) that govern these rare cells is a challenging issue [Gradilone et al., 2011, Giuliano et al., 2014].

1.2 Measuring gene expression using microarrays

DNA microarrays are a high throughput technology used to measure the expression levels of thousands of genes, in some cases all of the genes in a genome, simultaneously. The fundamental idea behind most microarrays is to measure the amount of the different types of mRNA molecules in a cell, thus indirectly measuring the expression levels of the genes that are responsible

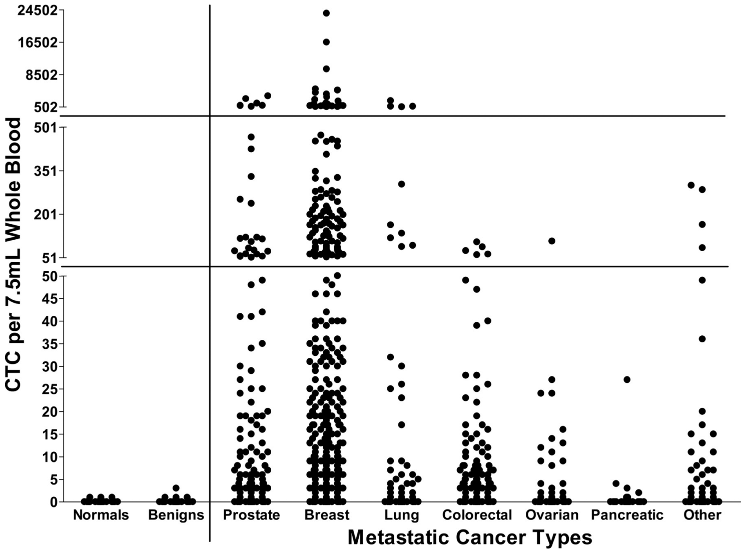


Figure 1.7: Prevalence of CTCs in 7.5 mL of blood of 145 women donors, 199 women with nonmalignant diseases, 188 samples from 123 metastatic prostate cancer patients, 1,316 samples from 422 metastatic breast cancer patients, 168 samples from 99 metastatic lung cancer patients, 333 samples from 196 metastatic colorectal cancer patients, 53 samples from 29 metastatic ovarian cancer patients, 21 samples from 16 metastatic pancreatic cancer patients, and 104 samples from 79 patients with other metastatic cancers.(Reprinted with permision from [Allard et al., 2004].)

for the synthesis of those particular mRNA molecules [Brown and Botstein, 1999, Lockhart and Winzeler, 2000].

Depending on the type of the probe material, microarrays can be employed for different purposes, e.g. transcriptomics (DNA microarray) or proteomics (Protein microarray) data analysis. In this thesis, the term “microarray” will mostly be used to refer to DNA gene expression microarrays, rather than protein or other types of arrays.

The concept behind this technology relies on accurate binding, also called hybridization, of strands of DNA with their precise complementary copies in experimental conditions where one sequence is also bound onto a solid state substrate (glass). Basically, a microarray chip is composed of DNA fragments (probes) that represent specific gene coding regions. Purified RNA fragments from a biological sample are then fluorescently or radioactively labeled and hybridized to the chip. Once the hybridization is complete, the chip is washed to remove non-hybridized fragments. The chip is then processed by a laser scanner in order to detect the areas of the chip where hybridizations occurred (see Figure 1.8).

Microarrays can be used to measure either absolute transcript concentrations or relative transcript concentrations (i.e. expression ratios). Traditionally,

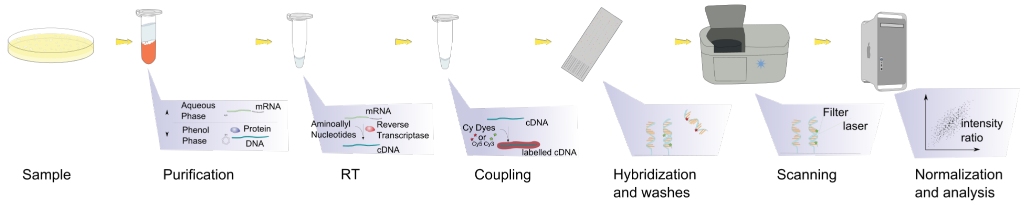


Figure 1.8: The steps required in a microarray experiment. Original work available in Wikipedia

two-channel, cDNA array data (e.g. using Cy3 and Cy5 dyes) are usually used to measure ratios, whereas single channel, oligonucleotide array data (e.g. Affymetrix) are intended to represent absolute expression values.

1.2.0.1 The Affymetrix GeneChip®

The GeneChip®, which is manufactured by Affymetrix, is an oligonucleotide array and is the most commonly used type of DNA microarray. They differ slightly in operation from other kinds of arrays. Each array will contain hundreds of thousands of probe spots and each of these spots will in turn contain millions of copies of the individual 25 base long DNA oligonucleotide.

Each gene that is being targeted is represented by typically (but not necessarily always) 11 pairs of these probes [Irizarry et al., 2003]. This set of probes contains 11 perfect match (PM) probes, which are exactly complementary to the DNA sequence of a subset of 25 bases of the target gene. Each PM probe has a corresponding mismatch probe (MM), which contains the same 25 base long sequence as the PM probe, except for the fact that the middle base, or the 13th base in the chain, is substituted for the complement of the 13th base of its corresponding PM probe; so for example, a G in the 13th base of a PM probe will be replaced with a C in the MM probe. This is meant to give an estimate of non-specific binding, which occurs when mRNA that is not targeted binds to a PM probe.

1.3 Analysis of gene expression data

The typical input of the computational stage in an experimental study using microarrays is the expression values of multiple (usually thousands) of genes measured in a number of different conditions or experimental setups. The different conditions could be different individuals that can have their own characteristics and phenotypes (e.g. clinical information, cancer stage, treatment plans, etc), or could correspond to different time points. There are public databases like the Gene Expression Omnibus (GEO[3]) [Edgar et al., 2002a] and ArrayExpress[4] [Parkinson et al., 2009] that store data from high-throughput functional genomic experiments, and provides these data for reuse to the research community.

Genotype and phenotype

The *genotype* is the unique genome of an individual and therefore corresponds to the individual's complete heritable genetic information.
The *phenotype* is a description of the actual physical characteristics. This can include a person's appearance, like eye color and height, the disease status and history, etc.

The data for the downstream statistical analysis are usually presented in matrix form $X \in \mathbb{R}^{M x N}$, where M is the number of genes (rows) and N is the number of conditions or observations (columns). This matrix is usually referred to as the *gene expression matrix*. The values $x_{i,j}$ in this matrix are numeric, denoting the absolute or relative intensity (expression) of a gene i in a condition j.

The data at this point may need to undergo a certain range of preprocessing steps. Typical preprocessing techniques include *imputation* in order to address missing values, *outlier detection*, and removal of features that appear to be redundant. Missing data can occur systematically as a result of the microarray fabrication, especially for cDNA arrays, or for various reasons such as insufficient resolution on images or due to dust and scratches in the slides. Filling the missing data with techniques such as the "K-nearest neighbors" [Troyanskaya et al., 2001] can address this problem and can produce complete data matrices for downstream analyses, although many alternative approaches exist [Liew et al., 2011, Shashirekha and Wani, 2015]. Outliers are measurements that are substantially different from the majority of the other values and can severely effect further statistical analysis. In order to identify them, Z-scores or coefficients of variation, and their robust versions using the median and the median absolute deviation (MAD), can be used with a cutoff value[5]. The values exceeding the cutoff threshold can then be replaced with the maximum permitted value ("trimming"). Finally, features that do not exhibit much variation can be totally excluded in order to reduce the number of genes to evaluate. To this end, measures of variability such as the standard

[3] http://www.ncbi.nlm.nih.gov/geo/ (accessed on May 6, 2016)

[4] https://www.ebi.ac.uk/arrayexpress/ (accessed on May 6, 2016)

[5] The z-score of a gene g that has mean value $\hat{g}$, standard deviation σ, and median $\tilde{g}$ is $(g_j - \hat{g})/\sigma$. The coefficient of variation is defined as $\sigma/\hat{g}$ while its median absolute deviation is the median value of $|g - \tilde{g}|$

deviation or the “interquartile range” are used[6].

Typical preprocessing steps may also include specific data transformation, such as “standardization”, logarithmic (log) transformation, and discretization. Standardization or normalization attempts through a linear transformation to make all genes have the same range of values, so that the statistical or data mining tools used next are not affected by features that show larger 1st order (e.g. mean) or 2nd order (e.g. standard deviation) statistics. Centering the values by subtracting the sample mean, median, or minimum value, and standardizing by dividing by the standard deviation (z-scoring), MAD, or the range (min-max normalization) are used frequently. The log transformation, instead, is a non-linear transformation and it is very popular for a lot of reasons. First, the expression values usually exhibit high skewness, with a lot of genes having low intensities followed by a long tail of genes that show larger intensities. The logarithmic transformation reduces much of this skewness, making the visual inspection of the data easier. It also improves the estimation of the variance and make the data amenable to statistical methods that assume normality. Other transformations, such as Tukey’s proposed “ladder of power transformations”, can also be used [Box and Cox, 1964]. On the other hand, discretization allows the transformation of continuous data to discrete or even binary values since researchers are usually interested in identifying three categories of genes with respect to the increase or decrease in their expression when compared to a reference or to their mean value: (i) up-regulated (increased expression, enrichment, and transcript concentration), (ii) down-regulated (decreased expression), and (iii) non-regulated. A popular, non parametric discretization method is that of Fayyad based on Minimum Description Length Principle (MDLP) [Fayyad and Irani, 1993], but several others exist [Potamias et al., 2004, Pensa et al., 2004, Bolón-Canedo et al., 2010, Li et al., 2010].

After this preprocessing stage, the core of the statistical and computational analysis is performed. This analysis potentially includes the uncovering of *differentially expressed genes*, the identification of hidden *clusters* in the data, or the generation of hypotheses and models for the *classification* of new unseen data into a set of known categories. We describe each of these tasks in the following paragraphs.

1.3.1 Differentially expressed genes

Researchers and biologists are especially interested in identifying genes that behave differently in different conditions than the majority of the other genes.

[6] The interquartile range is defined as the difference between the 25% and 75% quartiles. By definition, this range contains the 50% of the observed values.

When we analyse gene expression data from microarrays, the different behavior is related to the amount of the mRNA produced and therefore to the intensity values in the gene expression matrix. We call these genes "differentially expressed" [Dudoit et al., 2002b].

Finding the differentially expressed genes is not important only for the biologists but it is usually crucial for the statistical analysis as well. As we describe in more detail in Section 1.4.1, the size difference between the genes (rows) and the conditions (columns) in the input data puts a lot of strain to the subsequent data mining and machine learning processing. The identification of a small subset of informative genes will then greatly alleviate this problem.

Usually, the motivation behind the use of microarrays is a research question involving the comparison of two groups or conditions, such as the cancer samples versus the healthy ones, although more classes are also frequently compared. In such comparative two-group experiments the simpler approach is to check the expression of each single gene in the two groups, while more complex approaches could consider the expression of many genes in combination. Visualization techniques such as "heatmaps" (Figure 1.9), boxplots, etc. can be used as a first step to get a coarse understanding of the expression of genes in the different conditions. Nevertheless, since we try to use the expression of genes in a limited number of cases in order to draw general conclusions for the population, the analysis of gene expression data lies in the realm of the "statistical inference" [Casella and Berger, 2001]. There is a need therefore for quantification, and the calculation of the statistical significance of any findings is quite important.

One of the most simple strategies for finding differential expression of genes is the so called *fold change*, where a gene was considered to have significantly different expression in the two conditions or groups if the ratio of its mean expressions in the two groups is above some cut-off value. Usually, the comparison is done in the logarithmic scale and therefore a two-fold change (i.e. double of the expression) means that the difference of the mean expressions to be 1. The Fisher and Golub scores put similar focus on the difference of a gene's average expression in the two conditions, but also emphasize their "separability" by incroporating the standard deviations, as shown below [Golub et al., 1999]:

$$S_{\text{Fisher}} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \qquad S_{\text{Golub}} = \frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2} \tag{1.1}$$

where μ_1, μ_2 are the mean expression values of the gene in the two conditions and σ_1, σ_2 the corresponding standard deviations. Based on these scores the genes can be ranked in terms of their discriminating power or separability of the two classes or conditions.

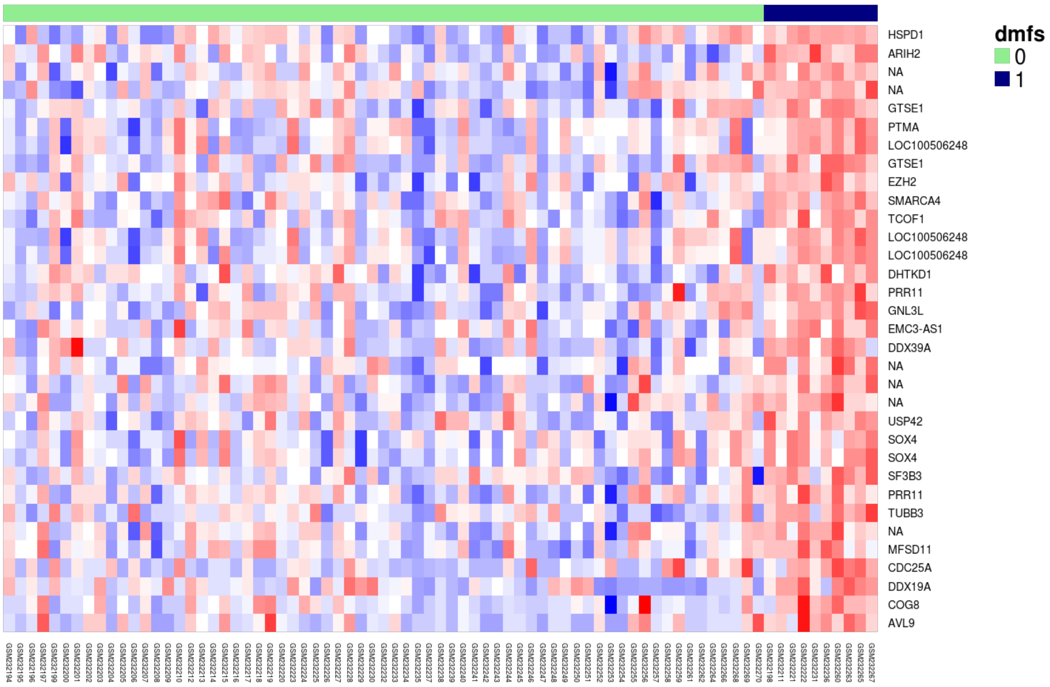


Figure 1.9: An example of a "heatmap" visualization, from [Notas et al., 2015]. Each column represents a sample and each row is a gene. Samples have been grouped according to the "dmfs" ("distant metastasis, free survival") variable. The selected genes are shown to have different expression in the metastatic samples.

Similarly, the two-sample *t test* is a generalization of the fold change that takes into account also the variability (standard deviation) of the gene's expression in the two conditions. *t* tests have been used extensively[7] due to their flexibility (e.g. one can assume equal or different group variances (i.e *Welch's test*)) and easy computation, and their interpretability and the strong connection with the *null hypothesis significance testing* (and the notorious p-value). Nevertheless, because of their sensitivity when the variance is low, modified versions of *t*-test have been proposed such as the "moderated t statistic" of [Smyth, 2005] and the *Significance Analysis of Microarrays* (SAM) [Tusher et al., 2001]. SAM in particular uses the following adjusted *t* statistic:

$$d(g) = \frac{\bar{x}_1(g) - \bar{x}_2(g)}{s(g) + s_0} \tag{1.2}$$

where $\bar{x}_1(g)$, $\bar{x}_2(g)$ are the mean expression values of gene g in the two conditions, $s(g)$ is its "gene-specific scatter" (pooled standard deviation), and s_0 is a "fudge factor" computed so that it minimizes the coefficient of variation

[7] Perhaps even excessively!

of the $s(g)$ values (i.e. for all genes)[8]. SAM then performs a large number of permutations, computes $d_p(g)$ for each permutation p, and averages these values to estimate the "expected relative difference" $d_E(g)$. A user specified Δ parameter can be used to identify genes that deviate more than that from their expected relative difference and consider them as significantly differentiating genes. At the same time, this parameter can estimate the number of "false discoveries" based on the permutations, and therefore the False Discovery Rate (FDR, see box below) is automatically calculated. In reverse, a researcher can pre-specify a cut-off value for the FDR, i.e. the expected proportion of false discoveries they can tolerate, and then SAM can calculate and use the corresponding Δ parameter.

False-discovery rate (FDR) and q-value

The False Discovery Rate (FDR) is the expected proportion of false discoveries (type 1 errors) among all statistically significant hypotheses identified by a hypothesis test [Benjamini and Hochberg, 1995a]. It is used to adjust a hypothesis test in a multiple testing scenario. Given a number of false positives V and a number of all positives R, the FDR can be written as: $\text{FDR} = \mathbb{E}(V/R)$. The q-value for a specific hypothesis test is the minimum FDR threshold at which the test would be regarded as significant.

There are a lot more techniques, ranging from model based methods, e.g. ANOVA (Analysis of Variance) models [Kerr and Churchill, 2001] to non-parametric tests such as RankProd [Hong et al., 2006]. [Drăghici, 2011] contains an extensive list of methods and their background.

1.3.2 Cluster Analysis

When the research question is related to the identification of sets of genes or sets of experiments that exhibit common behavior (e.g. similar gene expression patterns), a large assortment of techniques for *clustering* are relevant. The goal is to discover these *clusters* (groups) of observations in order to reveal potentially new phenotypes when clustering the observations (samples, experiments), or to identify genes with similar functions. In the machine learning community this is usually called *unsupervised classification* since there is no apriori knowledge about these groupings.

[8] In more detail, [Tusher et al., 2001] propose the following computation: Let s^{α} be the α-th percentile of the $\{s(g)\}$ values, and $d^{\alpha}(g)$ the gene specific value of Equation (1.2) when it uses s^{α} as the fudge factor. For $\alpha \in \{0, 5, \ldots 100\}$, compute the median absolute deviation (mad) from the median, $\text{mad}_j(\alpha)$, for each $[q_j, q_{j+1}]$ interval of percentiles. Then, find the $\hat{\alpha}$ value of α that minimizes the coefficient of variation of these $\text{mad}_j(\alpha)$ values and use the corresponding $s^{\hat{\alpha}}$ as the fudge factor.

Cluster analysis techniques can be hierarchical, when the result is a hierarchy of virtually nested groups, or partitional (non-hierarchical). *Hierarchical clustering* can be divisive, i.e. top-down, or agglomerative, i.e. bottom-up. In either case, the choice of the distance metric and the “linkage” method to be used is important. Typical choices for distances include the Euclidean or the (inverse) correlation coefficient[9]. The linkage criterion defines the distance between two groups of objects based on the distances of their members and it’s used for the splitting of groups, in the case of divisive clustering, or the merge of groups in the case of agglomerative. Possible cases for the linkage used are the “single” linkage (the minimum pairwise distance between the objects of the different sets), the “complete linkage” (the maximum pairwise distance), the average linkage, and the Ward’s criterion that splits or merges the groups in order to minimise the total within-group variance. Hierarchical agglomerative clustering has been used extensively in bioinformatics after the seminal paper of [Eisen et al., 1998].

In the partitional or non-hierarchical clustering, algorithms try to find a division of the set of data points into non-overlapping clusters such that each data point is in exactly one cluster. Example of a non-hierarchical algorithm is the k-means where each cluster is represented by its “centroid” [Hartigan, 1975]. In k-means, each data point is assigned to the cluster with the “closest”, in terms of Euclidean distance, centroid. Another example is the *Partitioning Around Medoids* (PAM) that implements the k-medoids algorithm [Kaufman and Rousseeuw, 2009]. Similarly to k-means, this algorithm attempts to find a segmentation of the input data so as to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. The difference is that k-medoids choose a specific data point, called “medoid”, as the cluster center in each iteration and can use non-Euclidean distances as well. PAM is a common realization of the k-medoids clustering that uses greedy search in order to find the solution: at each iteration a swap is performed between a medoid object i and non-medoid object j if this re-arrangement produces a better clustering. Another adaptation of PAM uses a *silhouette measure*, which contrasts the average proximity of a data point to the other points in the same cluster with its average proximity to the data points in the nearest cluster to which it is not assigned [Van der Laan et al., 2003]. Finally, model based clustering [Yeung et al., 2001], such as Gaussian mixture models, can provide a well-grounded, theoretically appealing statistical model for the evaluation of clustering results, but some structure needs to be imposed to limit the number of parameters to be estimated [Banfield and Raftery, 1993] (see also Section 1.4.1.1 for the regularization techniques).

[9] If ρ_{ij} is the Pearson correlation coefficient of two vectors i and j, then $d_R \equiv 1 - \rho_{ij}$ can be used as distance metric since similar (e.g. linear dependent) vectors will have $d_R = 0$.

In the partitioning class of cluster algorithms the number of clusters to identify is given as input. The problem of how many clusters to search for can be usually addressed by techniques such as cross validation or resampling where increasing values for the number of clusters are examined and the one that maximizes a criterion, like the average silhouette, is selected. A notable resampling technique is *consensus clustering* where different values for the number of clusters are tried and the most "stable" clustering is chosen [Monti et al., 2003].

Finally, more "exotic" cluster analysis techniques include biclustring and spectral clustering. Biclustering is a set of techniques that perform clustering in the two dimensions (samples, genes) concurrently [Madeira and Oliveira, 2004, Alevyzaki et al., 2016]. Contrary to the previous methods, biclustering algorithms try to identify groups of genes that show similar activity patterns under a specific subset of the experimental conditions or groups of samples. Non-negative matrix factorization is a linear factorization that can be used for biclustering [Carmona-Saez et al., 2006]. Spectral clustering, on the other hand, uses the spectral decomposition of an input similarity matrix to derive the first $\mathtt{k}$ eigenvectors and then performs clustering using $\mathtt{k}$-means in the matrix containing these eigenvectors. More details for spectral clustering can be found in [Pentney and Meila, 2005].

1.3.3 Classification

The landmark paper of Golub et al [Golub et al., 1999] represented the first demonstration gene expression profiling could be used to identify new cancer subtypes or assign tumors to known classes. Interestingly, the paper also demonstrated the use of unsupervised learning (clustering) for a supervised learning task.

[Dudoit et al., 2002a] and [Rogers et al., 2005] provide a relevant review of techniques for supervised learning but there are many classification algorithms originated from the Machine Learning community [Murphy, 2012]. [Slawski et al., 2008] provides a comprehensive list of such techniques used in the microarray data analysis.

1.4 Challenges in the microarray analysis

At first sight, microarray data analysis can be framed as an application of the classical statistical analysis of experimental data as formulated by the pioneering works of Karl Pearson, Ronald A. Fisher, Jerzy Neyman and others, but this could not be further from truth. There are a couple of issues innate to the application domain and the technologies used that make the analysis of microarray data quite challenging. In this section, we describe in more detail some of these issues.

1.4.1 High Dimensional Data

The data analysis with microarray gene expression data is always challenging because of the high dimensionality inherent in these datasets. For example it is frequently the case that we have a few tens of samples/cases and on the other hand thousands of genes to be used as features. The high dimensions, not only in bioinformatics but also in image processing, text mining, and other fields, are characterized by unintuitive geometric properties that not only puzzle the researchers but also put familiar 2-d or 3-d computational methods in a lot of trouble.

Pedro Domingos has provided a great summary of the problems one face in higher dimensions [Domingos, 2012]:

> Our intuitions, which come from a three-dimensional world, often do not apply in high-dimensional ones. In high dimensions, most of the mass of a multivariate Gaussian distribution is not near the mean, but in an increasingly distant "shell" around it; and most of the volume of a high-dimensional orange is in the skin, not the pulp. If a constant number of examples is distributed uniformly in a high-dimensional hypercube, beyond some dimensionality most examples are closer to a face of the hypercube than to their nearest neighbor. And if we approximate a hypersphere by inscribing it in a hypercube, in high dimensions almost all the volume of the hypercube is outside the hypersphere. This is bad news for machine learning, where shapes of one type are often approximated by shapes of another.

[Hayes, 2011] provides an accessible description of these "strange" geometric properties in higher dimensional spaces using the behavior of a unit ball as an example. The volume of a ball of radius r in the n dimensions is given by the formula [Scott, 2015]:

$$V_n(r) = \frac{\pi^{n/2} r^n}{\Gamma(1 + \frac{n}{2})} \tag{1.3}$$

where Γ is the Gamma function, for which we have $\Gamma(x+1) = x\Gamma(x)$ and in the case of natural numbers $\Gamma(m+1) = m!$. It can be easily shown that by increasing the number of dimensions by two, that is going from n to $n+2$, the

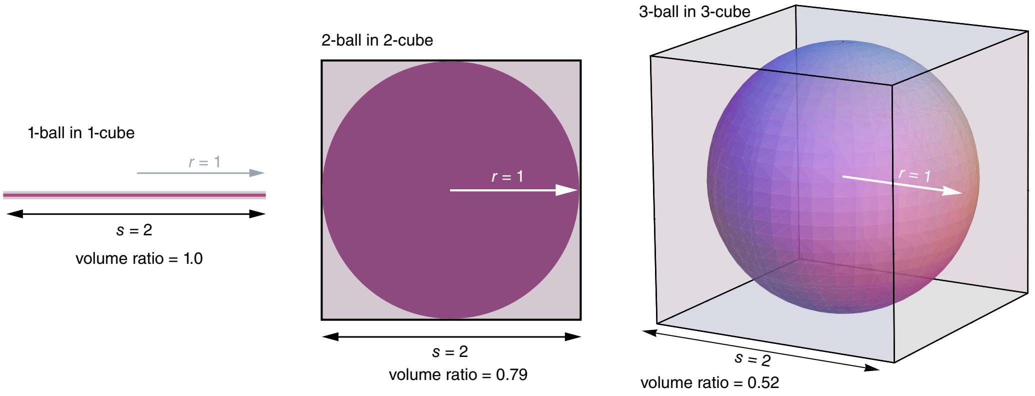


Figure 1.10: A unit ball is inscribed in the center of a "cube" in different dimensions. The cube has sides of length 2, which makes it just large enough to accommodate a ball of radius 1. In one dimension (left) the ball and the cube have the same shape: a line segment of length 2. As dimension increases, the ball fills a smaller and smaller fraction of the cube's internal volume. In three dimensions the filled fraction is about half. (Reprinted with permission from Brian Hayes and American Scientist [Hayes, 2011])

volume of the unit ball (i.e. with $r = 1$) actually decreases for $n \geqslant 5$, because the following recursive formula holds:

$$V_{n+2}(1) = \frac{2\pi}{n+2} V_n(1) \tag{1.4}$$

Therefore for $n \geqslant 5$ the factor $\frac{2\pi}{n+2}$ becomes less than 1 and the volume of the ball becomes smaller and smaller. The implications of this rather bizarre phenomenon are that if we assume a corresponding "cube" with sides of length 2 that can be put around the ball so that the ball touches its sides, in higher dimensions the volume of the ball becomes so small that the ball effectively vanishes, despite the fact that it is actually the "largest" sphere that be inscribed to it (see Figure 1.10). This is also called the *empty space phenomenon* [Scott and Thompson, 1983].

Another implication of this is that if we assume a large number of points uniformly distributed in the multidimensional space then as we go in larger dimensions the points go far apart from each other. This is the usual interpretation for the term "curse of dimensionality": in high dimensions, almost all pairs of points are equally far away from one another. To see why this is the case, imagine a single point x_0 as the center of our unit ball and N additional points scattered uniformly around it inside the inscribing cube, so that in each coordinate they are at distance 1. The volume of the ball corresponds to the average number of points that are at most at (Euclidean) distance $r = 1$ from the x_0 point. In higher dimensions the volume of the cube increases to be 2^n while, as explained above, the volume of the ball given by Equation (1.3) shrinks constantly so that effectively the space becomes too sparse and none of the N points are in the neighborhood of x_0. Where did all the points go? If not inside the ball, they should be in the “corners” of the cube!

Also if we imagine two “concentric” balls, one embedded to the other, with the ratio of their radii to be 0.9[10] then according to Equation (1.3) the ratio of their volumes decreases exponentially with rate 0.9^n when we consider an n-dimensional space. So in the 1-d case the “balls’ volumes” differ by 10% but when we consider the 5 dimensional space the relative difference of their volumes decreases to 40% while in the 100-d space the small ball has effectively vanished inside the bigger one.

The high dimensionality has strong implications also in the nearest neighbor types of algorithms. As [Beyer et al., 1999] proves, the minimum and the maximum distance between a random reference point Q and a list of random data points becomes indistinguishable compared to the minimum distance, i.e.

$$\lim_{d\to\infty}\left(\frac{\mathrm{Dist}_{\max}^{(d)}-\mathrm{Dist}_{\min}^{(d)}}{\mathrm{Dist}_{\min}^{(d)}}\right)\to 0$$

when $\lim_{d\to\infty}\mathrm{var}\left(\frac{||X_d||}{E(||X_d||)}\right)=0$[11]. This means under the given assumption that the relative contrast between near and far neighbors diminishes as the dimensionality increases. This property has similar effects to any technique that uses some notion of geometric distance such as the (Gausian) Radial Basis Function (RBF) kernel.

Additionally, many statistical techniques and machine learning methods assume that the number of cases N far exceeds the number of features d. When that’s not true (i.e. when $d \gg N$) problems like collinearity and numerical instability appear and result in failures of the methods. Furthermore, in such cases, most learning algorithms “overfit” the training data, i.e. one can easily find a decision function that separates the (sparse, due to the high dimensionality) training data set, but performs poorly on new, unseen data.

1.4.1.1 How to deal with the high dimensionality?

Theoretically, the challenges introduced by the high dimensionality can be resolved if we use a large number of points (samples, examples) that cover the whole space. But this is impractical because the number of points needed

[10] For example the bigger ball could be a unit ball and the other could have radius of 0.9.

[11] This assumption means that the ratio of the variance of the length of any point vector with the length of the mean point vector converges to zero when dimensionality increases, and holds for any L_p norm for $p \geqslant 1$.

increases exponentially with the dimension: if N points suffice to cover the feature space in the one dimensional (1D) case, in 2D we need N^2, N^3 in the 3D space, etc. Especially in the application domain that we study here, the bioinformatics analysis of gene expression data, the cost of acquiring new samples is very high and therefore the typical data set size is measured in tens or couple of hundreds of patient samples, in sharp contrast to the number of genes measured that is in the range of tens of thousands. In fact, it has been argued that thousands of samples are necessary in order to derive a robust list of genes for cancer prediction [Ein-Dor et al., 2006]. An approach to circumvent this problem could be to integrate multiple data sets in order to increase the number of samples/observations but this introduces its own array of challenges, which we expand in Section 1.4.3.

Therefore, an obvious way to tackle the problem is to try to reduce the dimensionality by performing "feature selection" or "variable ranking" [Guyon and Elisseeff, 2003] in a way that potentially removes irrelevant (non-informative) features. There are many different algorithms that use heuristics to select the "best" (or good enough) set of features (dimensions) since it is usually intractable to check exhaustively all possible combinations. Depending on the evaluation criteria used for the selection of a specific set of features, the methods are usually classified as "filters", "wrappers", or "embedded" [Saeys et al., 2007].

- In the filter techniques the features are evaluated by looking only at the intrinsic properties of the data, a score is calculated, and low-scoring features are removed. For the case of *univariate* filters, the relevance of the features can be estimated separately by using, for example, their linear correlation coefficients with a target variable of interest, or similar criteria like the *information gain* [Quinlan, 1986]. *Multivariate* filter methods are able to find relationships among the features but are less scalable and slower than the univariate filter techniques. A notable example of a multivariate filter method is the Correlation-based Feature Selection (CFS) [Hall, 1999]. Popular choices for filtering genes are also described in Section 1.3.1 where we discuss techniques (mostly univariate filters) for finding "differentially expressed" genes.

- The wrapper techniques, on the other hand, require one predetermined learning algorithm in feature selection and use its performance to evaluate and determine which features are selected. Examples of such techniques are the sequential forward selection, where features are added as long as the performance of the learning algorithm improves, the sequential backward elimination, where, in reverse, features are removed, genetic algorithms, etc.

- Finally, in the embedded methods, the search for the optimal subset of features is built into the learning algorithm, such as the Decision Tree family of algorithms, Random Forests [Breiman, 2001], or the Support Vector Machines (SVM) Recursive Feature Elimination (RFE) that uses the magnitude of the SVM weights as ranking criterion [Guyon et al., 2002]. As a subcategory of the wrappers-based feature selection, a range of methods incorporate feature "weighting" in their objective function, which are usually called *regularization* or *shrinkage* methods [Tibshirani et al., 2002]. Regularization is a technique that tries to reduce overfitting and variance by introducing additional bias and its origin dates back to Andrey Nikolayevich Tikhonov and his work on ill-posed problems[12] [Tikhonov, 1977]. Regularization, loosely speaking, means that while the desired classifier is constructed to approximately send the observed feature vectors to the correct labels, constraints are applied to the construction of the classifier with the goal of reducing the generalization error. Examples of regularization algorithms are the "Ridge Regression" (the classical Tikhonov regularization), the LASSO [Tibshirani, 1996], and the Elastic Net that combines both [Zou and Hastie, 2005]. In these methods, a penalization term is inserted in the objective function that is to be optimized, which results in certain uninformative features (dimensions) to be ignored (vanished) in the final solution. As an example, in LASSO regression, an l_1 norm is used in the objective function to penalize the coefficients of the features:

$$J(W) = \frac{1}{2N}\sum_{i=1}^{N}(w^T x_i - y_i)^2 + \lambda ||w||_1 \tag{1.5}$$

where $0 \leqslant \lambda \leqslant 1$ controls the amount of "shrinkage" of the parameters (the vector w of features coefficients in the regression) towards 0. For model based clustering or classification techniques using Gaussian distributions, shrinkage of the covariance matrices is a similar technique [Ledoit and Wolf, 2004, Schäfer and Strimmer, 2005].

Another way to fight against the curse of dimensionality is to try to locate a lower-dimension "embedding" of the data. Usually the data are not truly random but exhibit some "structure", making the *intrinsic dimensionality* a lot lower than the representational (*embedding*) one. For example, the surface of a sphere is a two-dimensional "manifold" wrapped around a three-dimensional object. In such cases, algorithms like Isomap [Tenenbaum et al.,

[12] Well-posed are problems that have a unique solution that depends on the data in some stable way. Ill-posed are problems that are not well posed.

2000], LLE [Roweis and Saul, 2000], Locality preserving projections [He and Niyogi, 2004], and other non-linear dimensionality reduction techniques [Lee and Verleysen, 2007] can be used effectively. A relevant array of techniques fall under the term *Topological Data Analysis* [Lum et al., 2013] that try to recognize shapes or patterns in the data, such as “loops” or linear segments, and then identify interesting groups using these shapes.

In other cases, a transformation of the original space can reveal hidden patterns where the data points are clustered together in areas of higher density. The most common such transformation is the Principal Component Analysis (PCA) that performs a linear decomposition of the original data set through the rotation and scaling of the axes in a way that most of the “total variance” in the original data set is persisted. Mathematically, the total variance equals to the trace of the covariance matrix (i.e. the sum of the variance in each dimension/feature) and using its “eigendecomposition” (or the Singular Value Decomposition (SVD) of the original data matrix, see Appendix A) we can derive a set of uncorrelated variables termed “principal components” (Equation A.7 in Appendix A). These principle components correspond to the eigenvectors of the covariance matrix and by keeping the first k of them we can approximate[13] as much as we want the original data matrix. So, for example, by keeping the first two or three principal components we can have a low-dimensional representations that covers a large degree of the variance (and thus a large degree of the information content) in the data, in spite of a potentially very high original dimensionality of the data. We can of course keep as many principal components (i.e. eigenvectors) as we want in order to decrease the reconstruction error and increase the total variance explained, at the expense of increasing the dimensionality in the transformed space. The so called “scree plot”, which shows as a decreasing function the variance explained by each principal components, can be used to decide how many principal components to keep [Cattell, 1966].

Similar techniques to PCA include the Multidimensional Scaling (MDS) and Sammon’s mapping [Borg and Groenen, 2005]. The input of MDS is a matrix consisting of the pairwise dissimilarities of the samples (observations), which are not necessary distances in the strict matthematical sense, and the algorithm tries to reproduce those dissimilarities in a reduced dimensional space. In the classical MDS (also known as Principal Coordinates Analysis), Euclidean distances are assumed and the method is the same as doing PCA in the matrix of distances. In the non-metric MDS the dissimilarities are known only by their rank order and therefore they are qualitative (e.g. ordinal), while the metric MDS is a generalization of the classical MDS where the dissimilarities are still

[13] According to the Frobenius norm, which is equal to the sum of of the squares of the singular values of the original data matrix [Meyer, 2000].

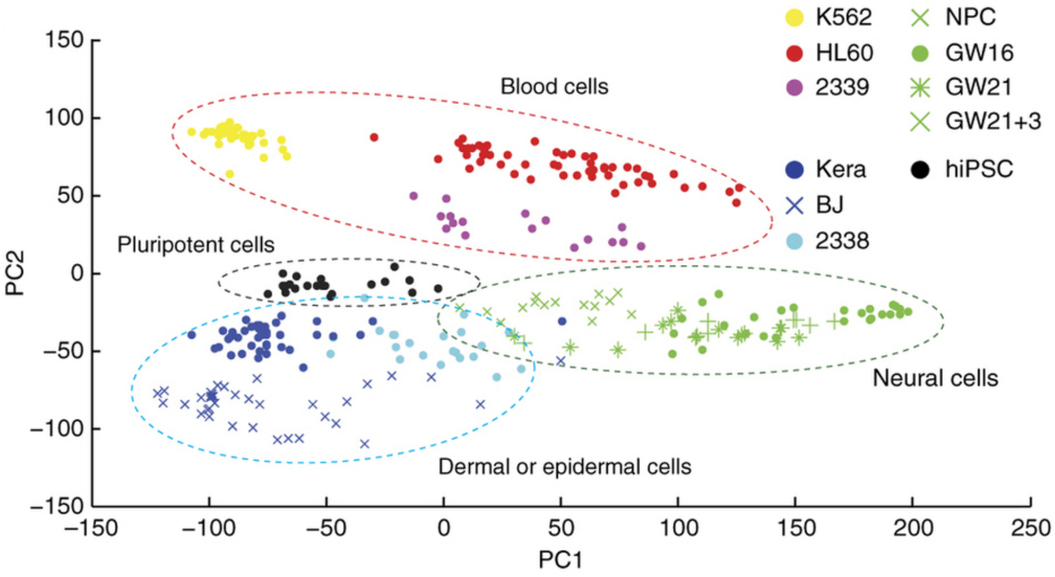


Figure 1.11: A "PCA plot", reprinted with permission from [Pollen et al., 2014].

quantitative but need not be Euclidean and in fact the optimization sought can take into account a parametric monotonic function of the original dissimilarities. On a related note, t-SNE is a non-linear dimensionality reduction technique that is targeted towards the 2D or 3D visualization of high dimensional data [Van der Maaten and Hinton, 2008]. All these techniques are valuable especially as a preprocessing step in order to gain insights on the data when visualized in two- or three-dimensional spaces. For example, in Figure 1.11 we see a 2D visualization of the high dimensional input data, revealing the existence of clusters that can be biologically explained.

When the dimensionality of the input data is very high the problem of scalability and efficiency of certain algorithms becomes evident. This is for example the case for the "instance-based learning" and Nearest Neighbor classifiers that use Euclidean or Mahalanobis[14] distances to determine the similarity or "nearness" of two points. To alleviate this issue, a celebrated lemma by Johnson and Lindenstrauss [Dasgupta and Gupta, 2003] asserts that a set of N points in high dimensional Euclidean space can be projected into a $O(\log N/\varepsilon^2)$ dimensional Euclidean space such that the distance between any two points changes only by a factor of $1 \pm \varepsilon$. Since Euclidean distances are preserved, running the Nearest Neighbor classifier on this mapped data yields the same results but at a lower computational cost.

[14] The Mahalanobis matrix is the inverse of the covariance matrix.

1.4.2 Batch Effects

[Leek et al., 2010] defines *batch effects* as "sub-groups of measurements that have qualitatively different behaviour across conditions and are unrelated to the biological or scientific variables in a study." The term batch denotes a collection of microarrays (or samples) processed at the same site over a short period of time using the same platform and under approximately identical conditions [Chen et al., 2011]. For example, batch effects may occur if a subset of experiments was run on Monday and another set on Tuesday, if two technicians were responsible for different subsets of the experiments, or if two different lots of reagents, chips or instruments were used. [Churchill, 2002] mentions some of the reasons for these effects:

> Slides are often printed in batches that can vary in their overall quality and even within a batch, the order and position on the printing device can affect results.

Usually the processing date of the microarrays is an important confounding variable that can reveal the presence of batch effects, as described in [Akey et al., 2007]. As an additional example, consider the data set of [Pau Ni et al., 2010] that is available from Gene Expression Omnibus (GEO) under the GSE15852 accession number[15]. The authors claim that they did find a differentiation between 43 breast carcinomas and 43 normal breast tissues collected from Kuala Lumpur Hospital, UKM Hospital and Putrajaya Hospital, in Malaysia. I wanted to use this dataset but usually the first step is to try to visualize the data if possible, in the two or three dimensions. An easy way to do this is to reduce the dimensionality by performing a Principal Component Analysis (PCA) or the corresponding Multidimensional Scaling (MDS), as described in Section 1.4.1.1, and keep the first two or three "principal conponents" as axes. The results in the 2-D case as shown in Figure 1.12 and reveal the existence of two "clusters" in the data, which do not seem to be related to the physiology of the samples, e.g. whether they come from healthy people or from cancer samples.

Hopefully, the data available from GEO contain also the date when the samples were processed and annotating the samples with the processing date seems to fully explain the bimodal distribution of the data: as can be seen in Figure 1.13 the cluster shown in the right contains samples processed in 2007 while the left cluster groups mostly the ones processed in 2006! Therefore the processing date appears to be a strong differentiating factor between the samples, although we don't know whether some third factor (e.g. heredity,

[15] The data set can be found at `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15852` (accessed in February,19 2016)

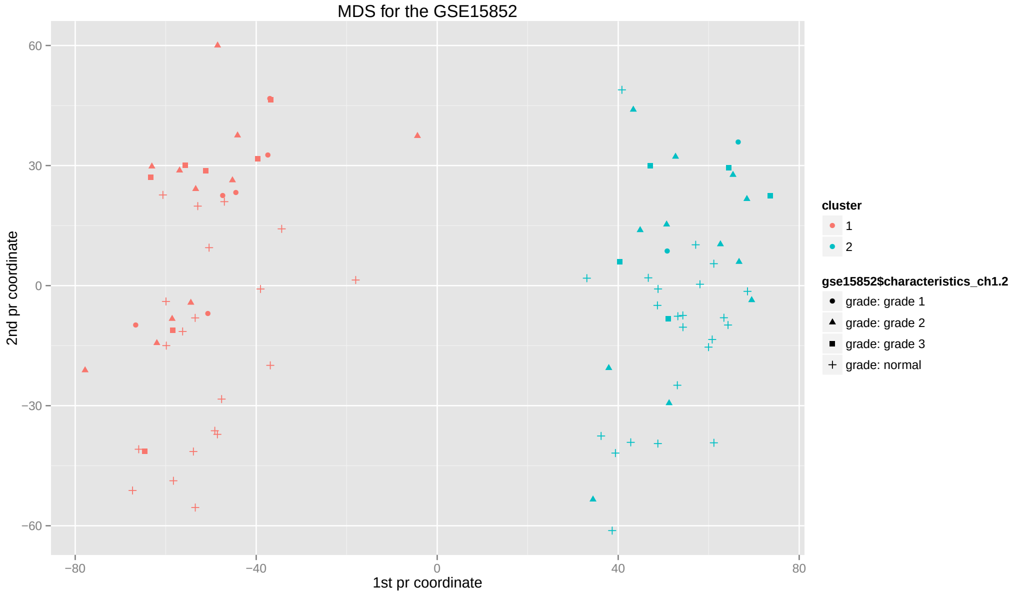


Figure 1.12: A PCA plot for the GSE15852 dataset reveals two groups of samples. The normal and cancer samples seem to be intermixed in both clusters and therefore there does not seem to be a correlation between the disease status and the cluster the samples belong to.

environment, etc.) can possibly "explain" both the assortment of the samples in the PCA plot and the processing date.

[Chen et al., 2011] provides an evaluation of six algorithms that attempt to remove batch effects. The "Empirical Bayes" or COMBAT algorithm [Johnson et al., 2007] appears to perform best. COMBAT estimates the mean and variance of each gene by pooling information from multiple genes with similar expression characteristics in each batch.

1.4.3 Integrative analysis of multiple data sets

For reasons related to the need to defend against the high dimensionality, to get better statistical power, and to validate or expand previous findings, integration of different microarray data sets appears to be an attractive approach [Rhodes and Chinnaiyan, 2005]. This is facilitated by the availability of freely accessible public repositories such as the Gene Expression Omnibus (GEO) [Edgar et al., 2002a] and the ArrayExpress [Parkinson et al., 2009].

Despite the benefits that this data integration can yield, using data from different microarrays studies is quite problematic [Tsiliki et al., 2011]. The danger of batch effects that we mentioned above (Section 1.4.2) is only one of

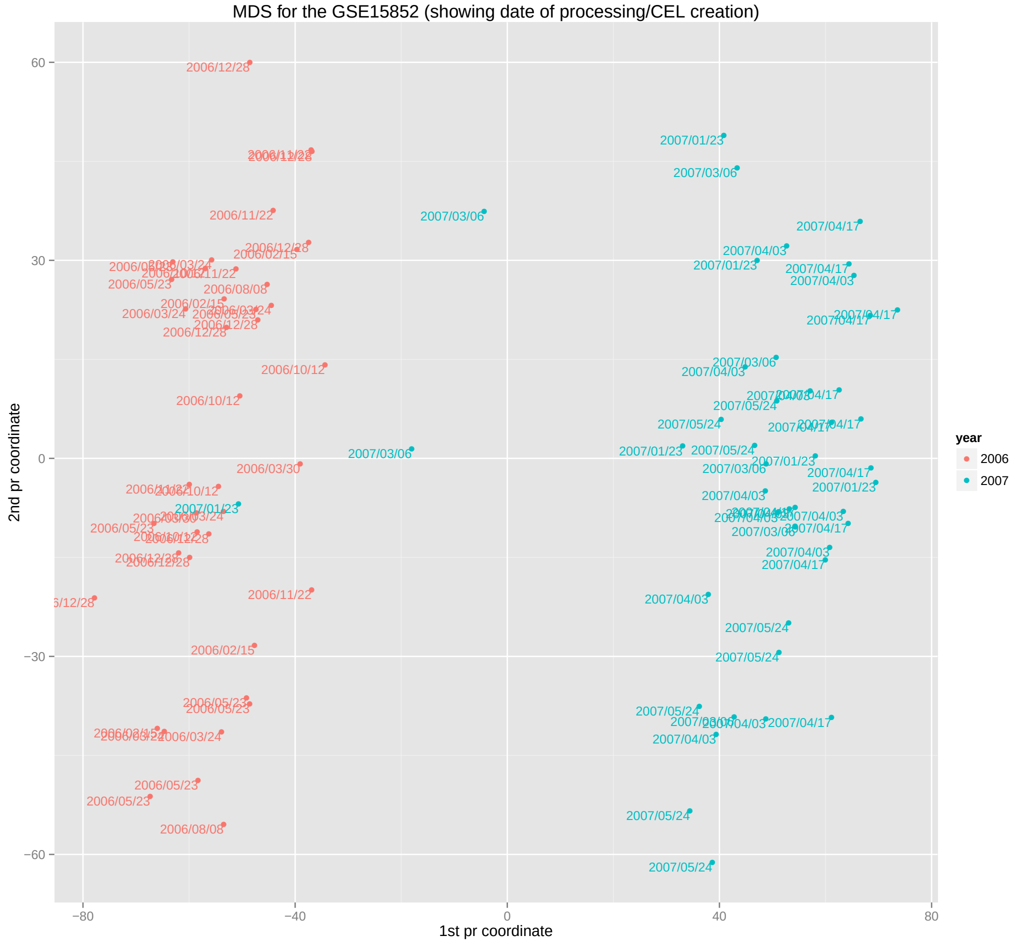


Figure 1.13: When we annotate the points with the date information i.e. when the corresponding microarrays were processed, we see that the right cluster contains samples processed in 2007 while the left cluster groups mostly the ones processed in 2006. The processing date is therefore a confounding factor for the gene expression variation between the samples.

the many challenges. Additional issues are the use of different platforms (e.g. cDNA versus oligonucleotide arrays), different probes (DNA transcripts), the varying length of those sequences, the lack of common annotation i.e. mappings from transcript identifiers to gene names and symbols, and the number and characteristics of the samples hybridized. In Section 2.2.2 we provide details on our strategy for addressing those problems.

1.5 Sources of biological knowledge: Biological networks and Gene Ontology

In response to the challenges imposed by the high dimensionality and the intricacies (e.g. noise) of the technologies used, the incorporation of expert prior knowledge aims to guide the knowledge discovery process or increase its performance and accuracy. An aspect of such an effort is the exploitation of existing knowledge as formulated in the contents of an ontology. An ontology describes "the objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold among them" and therefore ontologies have a very broad application domain and are more expressive than vocabularies, taxonomies, and thesauri. In the biological domain a number of ontologies have been developed in the recent years [Bard and Rhee, 2004], with Gene Ontology (GO) being the most popular. The GO is organized along three main axes, *molecular function*, *biological process*, and *cellular component*, where the concepts are represented though a taxonomy, going from the most general to the most specific terms [Ashburner et al., 2000]. Such a taxonomy can be easily represented as a graph. GO aims to represent our knowledge about biological processes, molecular functions, and cell components. In [Azuaje and Bodenreider, 2004] such semantic similarities suggested by GO were used to deduce correlation of gene expression data. [Bolshakova et al., 2006] suggest using the GO as the domain knowledge in order to validate clustering results and to determine the number of clusters in gene expression analysis. [Khatri and Drăghici, 2005] present a number of tools for GO based analysis of gene expression data.

Another source of domain knowledge comes from the various biological networks that represent complex reactions at the molecular level in living cells [Alm and Arkin, 2003, Alon, 2003]. The term Biological Networks has been introduced to describe any graph whose vertices are biological entities, such as genes, proteins, molecules, etc [Mason and Verwoerd, 2007]. With respect to the analysis of the gene expression data we can identify three major types of biological networks [Vidal et al., 2011]:

- Metabolic networks, which consist of chemical reactions that result in the construction of complex compounds and the storage and release of energy and are controlled by special proteins (enzymes). The nodes of these networks are biochemical metabolites and edges are either the reactions that convert one metabolite into another or the enzymes that catalyze these reactions. The edges can be directed or undirected, depending on whether a given reaction is reversible or not.

- Gene Regulatory networks that describe the relationships of genes, proteins, and other molecules with respect to the regulation of gene expression (i.e. the expression of a gene can be controlled by the product of another gene). These networks are directed graphs where the direction of an edge represents which gene or transcription factor affects the expression of another gene or regulatory element.

- Protein-Protein interaction networks (PPIs), where proteins (physically) interact with each other to promote the activation of a protein or to form protein complexes. The edges are undirected, as it cannot be said which protein binds the other, that is, which partner functionally influences the other.

The potential of the use of these networks and additional ones (such as disease-disease networks, i.e. graphs that show inter-disease connections [Goh et al., 2007], see Figure 1.14) has been eloquently argued in [Barabási et al., 2011], in addition to the current shortcomings, e.g. incompleteness, and "investigative biases". It's not surprising therefore that in the recent years biological networks have been the subject of important research: In [Rapaport et al., 2007] a general framework is presented that aims to analyse gene expression data when a gene network is known a priori; [Chuang et al., 2007] focus on identifying markers of metastasis within gene expression profiles where these markers are not encoded as individual genes or proteins, but as sub-networks of interacting proteins within a larger human protein–protein interaction network; The search for subnetworks of interaction networks that exhibit significant changes in gene expression has also been investigated in [Ideker et al., 2002]. In this work we make extensive use of Protein-Protein interaction networks and take advantage of their structure and their properties (e.g. degree distribution, random walks) in order to reveal underlying "neighborhoods" and get a better understanding of the genes biological functionality (Chapters 3 and 4).

1.5.1 Examples of using biological knowledge in computational methods

As descrined in Section 1.4.1.1, a popular approach is to use some lower reduction technique (PCA, SVD, etc) to find smaller number of "meta-genes" that are (linear, usually) independent. In [Chen and Wang, 2009] the authors classified the available genes using GO and for each gene category they constructed "super genes" by summarizing information from genes related to outcome using a modified principal component analysis (PCA) method. Then, they use these supergenes representing information from each gene category as predictors to predict survival outcome.

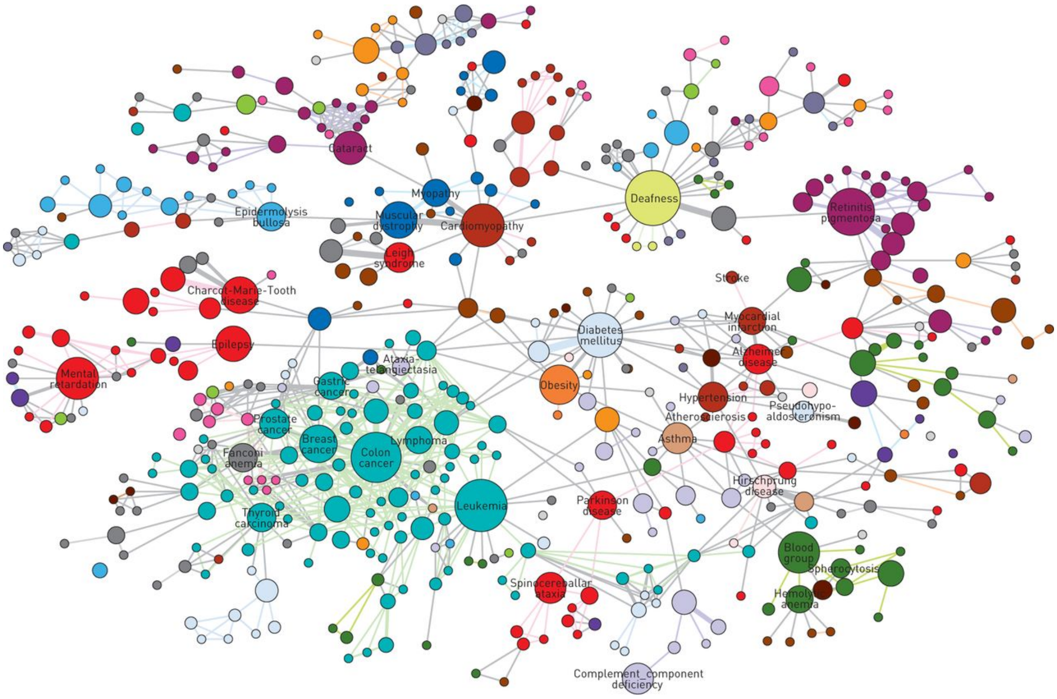


Figure 1.14: Graphical representation of the human disease network, where each node corresponds to a distinct disorder and colours represent disease classes [Goh et al., 2007] (Copyright (2007) National Academy of Sciences, USA). The size of each node is proportional to the number of genes participating in the corresponding disorder, and the thickness of the edge (link) is proportional to the number of genes shared by the disorders it connects.

Traditionally clustering requires a distance metric and below are some distance metrics that take advantage of the prior biological knowledge:

- [Hanisch et al., 2002] proposed the use metabolic networks for clustering tasks, by first defining a distance metric based on a pathway and then this metric is combined with the usual expression-based metric using their average. The combined metric is as follows ($o_k = (g_k, \nu_k)$ with g_k being genes and ν_k network nodes)

  $$\Delta(o_i, o_j) = 1 - 0.5(\lambda_{exp}(g_i, g_j) + \lambda_{net}(\nu_i, \nu_j))$$

  where

  $$\lambda_\Psi = \frac{1}{1 + e^{-s_\Psi(\delta_\Psi(x_i, x_j) - \nu_\Psi)}}$$

- In [Kustra and Zagdanski, 2006] the following distance metric is proposed:

  $$\text{dist}(g_1, g_2) = \lambda d(x_1, x_2) + (1 - \lambda) d(g_1, g_2)$$

where $d(x_1, x_2)$ is the distance between the gene expression profiles and λ is a user defined coefficient in the interval $[0, 1]$ that defines the balance between the expression and Gene Ontology based similarity.

- Huang and Pan [Huang and Pan, 2006] use the k-medoids algorithm (see Section 1.3.2) with a distance metric that is computed as $r \cdot d(x_1, x_2)$ if the two genes share a common functional annotation (e.g. participate to the same pathway) and $d(x_1, x_2)$ otherwise, with r being a shrinkage estimator in $[0, 1]$. The parameter r is estimated through cross-validation.

1.6 Objectives and design of the thesis

The prevalence of circulating tumor cells (CTCs) in peripheral blood of metastatic breast cancer patients has been evaluated by several groups and has been correlated with poor progression-free and overall survival of the patients, as described in section 1.1.3. CTCs are frequently found in the blood of patients with primary solid tumors, and it is generally assumed that a subset of these cells will eventually give rise to distant metastases [Labelle and Hynes, 2012]. Due to their involvement in the metastasis CTCs can indeed play an important role in the treatment of the cancer [Lianidou, 2014]. We can even potentially use them for testing the efficacy of existing tumour therapies like chemotherapy at the individualized level ("chemosensitivity") [Pachmann et al., 2014]. But first we need to find ways to detect CTCs in the blood of the patients, which is quite challenging in practice.

The technical challenge in this field consists of finding "rare" tumour cells and being able to distinguish them from epithelial non-tumour cells and leukocytes [Paterlini-Brechot and Benali, 2007]. The CTC frequency is usually low, often around 1-10 CTCs per millilitre of whole blood [Miller et al., 2009], which could also vary per disease and cancer progression (see Figure 1.7), while 1 millilitre of blood contains about 10 million white blood cells and 5 billion red blood cells [Barrett et al., 2010]. Numerous CTC detection techniques have been developed so far [Ignatiadis et al., 2015] but only a few of them have been used in clinical cohorts of relevant size. Among these technologies, some very promising approaches, such as the microfluidic-based "lab-on-a-chip", have not yet confirmed their better sensitivity in large cohorts [Nagrath et al., 2007]. In Breast Cancer patients, comparisons between the different techniques showed that filter-based and different EpCAM enrichment-based detection techniques have a globally similar CTC detection rate, although important differences were seen at the individual level [Magbanua et al., 2015]. Another approach to increase the CTC detection rate is to screen larger volumes of blood [Stoecklein et al., 2015], although higher numbers of CTC detected does not always translate into a better clinical validity or utility.

A major challenge in the CTC detection is the low concentration CTC in the blood, especially in post-operable breast cancer patients [Nadal et al., 2013]. Most current methods for detecting CTCs in patients score only single cells and could be missing an important fraction of the CTC population [Labelle and Hynes, 2012]. In lieu of these techniques, here we attempt to explore the molecular characteristics of the peripheral blood of breast cancer patients using statistical methods on the microarray data sets. This approach can be indirectly linked with the characterization of the circulating tumor cells, and it is based on the following assumptions and research findings:

- Distant metastases rely on the dissemination of tumor cells via the blood circulation, in a multi step process: detachment from the primary tumor, intravasation into the vascular system (whether directly or via the lymph nodes), survival while in transit through the circulation, extravasation and initial seeding for the creation of *micrometastases*, and finally the *colonization* in the distant organs, the proliferation, and the growth of macroscopic metastases [Labelle and Hynes, 2012]. Metastasis is an *inefficient process*, especially the latest steps, since the micrometastatic cells need to adapt to the microenvironment of the tissue in which they have landed, which is generally a lot different than the microenvironment of the tissue from which they originated [Chambers et al., 2002, Weinberg, 2007].

- CTCs carry information from primary tumor [Obermayr et al., 2010], but also from secondary tumor [Barbazán et al., 2012]. Moreover, [Sieuwerts et al., 2011] reported discrepancies in estrogen receptor and HER2 status profile compared to primary tumor, while it has been shown that metastases, which may develop several years after occurrence of the primary tumor and after prior systemic therapy in the adjuvant or neoadjuvant setting, can differ greatly from primary tumor tissue in terms of genetic characteristics [Suzuki and Tarin, 2007, Park et al., 2009].

- CTCs can be detected in single-cell level through specific genes. [Obermayr et al., 2010] identified six genes (CCNE2, DKFZp762E1312, EMP2, MAL2, PPIC and SLC6A8) as potential markers for the detection of circulating tumor cells in the peripheral blood of patients with breast cancer. Genes VIM, CXCR4, and uPAR were also found significantly correlated with the presence of CTCs by [Markiewicz et al., 2014].

- Cancer causes alterations in specific tissue areas but also in the blood. Gene alterations in tissue relate to those in blood, but also differences are important Specific gene alterations are (might be) indicative of the ability of cancer to diffuse in blood; such genes can predict the existence of CTCs without the need to exactly detect them [Molloy et al., 2012].

- CTCs diffuse in other blood cells and justify the analysis of blood volume percentage. Also, [Bettegowda et al., 2014] were able to detect the existence of circulating tumor DNA (ctDNA) (i.e. DNA of the dying tumor cells) in the blood of cancer patients even without the presence of CTCs. It therefore seems to be the case that specific differences of cancer tissue and cancer blood are indicative of the ability of tumor to diffuse and can be used for prognostic means and possibly for CTC assessment.

According to the above, a first objective of our work is to identify gene alterations in blood based on the characteristics of breast cancer tissues but also taking into account the specific blood characteristics so that only cancer-related markers are identified in the blood of the patients. In statistical terms, this means that any comparison between cancer tissue and cancer blood needs to *control* for any variation due to the origin (tissue or blood) of the samples.

Subsequently, any findings need to be linked with the prior biological knowledge in order to, if possible, explain and validate it, and potentially extend them to new discoveries. For example, it has long been discovered that despite the early successes (e.g. [Van't Veer et al., 2002], [Paik et al., 2004], [Wang et al., 2005]) a robust breast cancer gene signature has not yet been found [Ein-Dor et al., 2005, Weigelt et al., 2012]. However, integrating secondary data sources like protein-protein interaction (PPI) networks and other sources of biological knowledge has been proved to overcome the variability in the signatures and improve the predictive power [Chuang et al., 2007, Lee et al., 2008, Taylor et al., 2009, Abraham et al., 2010]. For these reasons, we are focusing on the use of biological networks for the validation and expansion of our findings.

1.6.1 Contributions and structure of the thesis

This thesis focuses on the characterization of circulating cancer cells in Breast Cancer patients using computational, data-driven, statistical methods. The goal is to identify differences and similarities between the blood and tissue samples of cancer patients and healthy populations using publicly available data sets that contain gene expression measurements. In order to proceed to a statistically sound genomic classification of tissue and blood of breast cancer patients a data integration approach has been designed. A large compendium of publicly available gene expression data sets has been brought together and carefully merged in order to overcome study specific biases or platform related technical variations. This integration methodology is then followed by a number of statistical comparisons between the different in origin (blood or tissue) or in disease status (cancerous or healthy) samples in order to reveal potential “biomarkers” for each case. These biomarkers are genes that exhibit different behavior (e.g. over-expression) in the aforementioned comparisons but in order to increase the sensitivity the sets of discriminating genes are intersected and a common subset is identified. The unique set of genes derived is then related to well curated sources of biological knowledge, such as biological networks, and subjected to novel algorithmic procedures so as to establish the underlying biological foundation and to further elicit features (genes) for the supervised and unsupervised classification of breast cancer patients.

The key contribution of this work is the discovery of a 27-genes signature as potential markers for the characterization of CTCs and the metastatic cascade, and a number of computational methods and their findings that take advantage of existing biological knowledge to fine tune the derived signature for the supervised or unsupervised classification of patient samples, as follows:

- Following the methodology described above, 9 different data sets publicly available from the Gene Expression Omnibus (GEO) database were assembled and integrated, yielding more than 800 samples of gene expression measurements. The subsequent statistical analysis and integration of results produced a “genes signature” of 27 genes as candidate biomarkers related to the presence of CTCs. In a subsequent, biological “bench-top” experiment, two of these genes (CXCR4 and JUNB) were in fact found to be really CTC-related, effectively confirming the statistical findings. The analysis that led to these findings is presented in detail in Chapter 2.

- In order to gain more insight in the induced signature of the 27 genes, we introduced prior domain knowledge in the form of biological networks. The question whether the genes participating in the derived signature are related and how they affect each other was formulated in as the graph theoretical problem called “Steiner Tree”. This formulation and the corresponding solution in a high quality protein-protein interaction network reveals the shortest interconnect for the genes in our signature and enhances it with additional central genes along the interconnecting paths. The methodology is based on the local properties of the genes in the graph, i.e. their immediate neighbors, the neighbors of their neighbors, and so on. Chapter 3 demonstrate these computational methods and their results.

- The incorporation of the prior biological knowledge in the form of biological networks using the induced graph of the genes in the vicinity of the 27 genes is described in Chapter 4. The goal is to take advantage of the "neighborhoods" of the 27 genes in an underlying biological network and to introduce a two-level classification scheme for new unseen samples. In contrast to the previous strategy, we now consider "random walks" in the graph, and therefore the approach is much more global and holistic.

- Finally, in Chapter 5, we present an adaptive model-based clustering that exploits prior biological information (e.g. groupings of genes) in order to perform "clustering" (unsupervised classification) of patient samples. The method effectively performs a biologically-inspired regularization in the well known "mixture of Gaussians" model. The method is presented first in the generic way and we then proceed to test the method when parameterized with the neighborhoods of the 27 genes.

The list of publications that were used for the dissemination of these results can be found in Appendix C.

Chapter 2

Computational Methods for the characterization of CTCs

Contents

In this chapter we present the main outcome of our work, which is a list of 27 genes that can potentially characterize the circulating tumor cells and the alterations of the genetic profile of the whole blood cells in the breast cancer patients. The approach is based on the computational data analysis and statistical tools using a large compendium of breast tissue and blood samples from patients and healthy subjects. The results are validated in independent datasets and their biological evaluation is presented.

2.1 Introduction

Nowadays, a large number of high-dimensional gene expression datasets are obtained through the exploitation of molecular techniques, such as DNA microarrays. Gene expression profiling of CTCs might provide the opportunity to identify markers for diagnosis and prognosis in breast cancer patients [Dirix et al., 2005], towards better provision of personalized medicine [Riethdorf and Pantel, 2010]. Furthermore, exploring gene alterations in CTC profiles could give valuable information on the molecular mechanism of tumor cell metastasis. In this chapter, we take advantage of CTC-targeted microarray studies obtained from human peripheral blood (PB) and tissue of breast cancer patients, as well as control individuals, in order to formulate a working hypothesis for the identification of a gene signature characterizing metastasis and the existence of CTCs.

2.1.1 Rationale and design methodology

The aim of this work is the identification of genes that characterize CTCs using comparisons between healthy (normal) and cancerous biological samples originated from tissues and blood. Our hypothesis supports that specific differences of cancer tissue and cancer blood are indicative of the ability of tumor to diffuse and, thus, can be used as factors for CTC estimation without direct detection. Direct gene expression profiling of CTCs is difficult for the reasons we explore in Section 1.6. Instead we aim at attacking the problem through an *indirect* approach, based on bioinformatics methods and separate datasets from blood and tissue samples.

For this purpose, we are using a two-stage procedure applied on several publicly available DNA microarray datasets from different origins (tissue and blood). The first stage aims to extract gene signatures associated with pair wise differentiation between cell types and /or disease states. For instance, the comparison of cancer and control tissue provides information about the discriminative factors of the primary disease. Next, the comparison between cancer blood and control blood can derive markers indicative of alterations due to the pathology, related to the CTC content and in association to the primary and secondary disease. From this stage, we derive four signatures, each reflecting the over-expressed differential profiles of genes associated with the specific comparison. In more detail we consider the following primary comparisons, which effectively test site-specific differences in the presence of the disease:

- **Cancer tissue versus healthy tissue** (C_1): What are the tissue specific differences during the disease progression? This comparison can reveal genes and biological processes that are triggered in the case of (primary) breast cancer patients.

- **Cancer blood versus healthy blood** (C_2): Are there any differences in the blood of cancer patients when compared to the blood of healthy subjects? The prevalent hypothesis for the relapse of the cancer patients and the appearance of metastases in distant organs is that tumor cells intravasate to blood circulation, and therefore we expect that this comparison provides strong evidence for the characterization of this process. Nevertheless, any identified differences in this comparison could be due to blood specific characteristics and cannot necessarily provide any CTC-related information.

We additionally consider the cross-site comparisons (blood - tissue) for similar differential expression in the presence of disease, as follows:

- **Cancer tissue versus healthy blood** (C_3): What are the genes that exhibit differential expression in cancer tissue when compared to normal blood? This comparison identifies genes over-expressed in primary cancer and not in blood cells and in combination with C_2 provides a fine tuning mechanism for the identification of cancer related differentiation in blood, as there is strong evidence that the metastatic tumors bear a lot of similarities with their primary cancer sites (e.g. [Ding et al., 2010]).
- **Cancer blood versus healthy tissue** (C_4): Likewise, this comparison further filters and specializes the identified differentially expressed genes to a set of potential biomarkers that exhibit elevated expression in the disease both in the blood and the tissue samples.

Figure 2.1 provides a graphical representation of this design. The hypothesis at this point is that peripheral blood from cancer patients carries information regarding the primary and secondary (metastasis) tumor, as well as other cancer induced alterations. By comparing the previous signatures, we can isolate markers indicative of certain aspects of cancer, leading to closer association with the existence of CTCs. This is the rationale for the second stage of our proposed procedure, which considers the intersection of the previous signatures. In particular, we consider the intersections:

- $\mathsf{C}_1 \cap \mathsf{C}_2$: derives genes over-expressed in cancer tissue and blood; however, it can also reflect genes over-expressed in normal blood,
- $\mathsf{C}_1 \cap \mathsf{C}_2 \cap \mathsf{C}_3$: eliminates genes over-expressed in normal blood and involves only genes over-expressed in blood due to cancer-associated factors.

- $C_1 \cap C_2 \cap C_4$: eliminates genes over-expressed in normal blood and involves only genes over-expressed in blood due to cancer-associated factors.

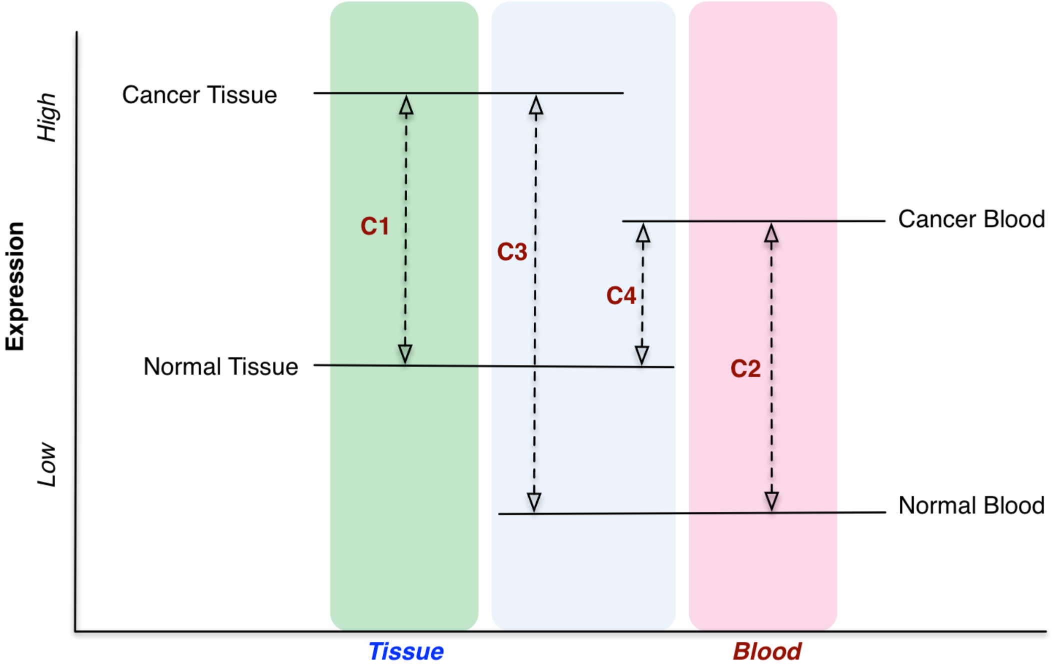


Figure 2.1: The setting we aim to find differentially expressed genes for. The different comparisons aim at finding over-expressed genes as shown in the figure: C_1 *compares tissue samples and yields the genes over expressed in Cancer;* C_2 *compares blood samples and provides the genes over expressed in Cancer;* C_3 *determines the genes over expressed in Cancer Tissue with respect to Normal Blood; and finally* C_4 *yields the genes over expressed in Cancer Blood when compared to Normal Tissue. Therefore* C_1 *and* C_2 *are comparisons in homogeneous samples (tissue and blood respectively), while* C_3 *and* C_4 *are both between blood and tissue samples. Since no comparison between the cancer tissue and cancer blood is performed, their relative position is not important and the image shows one possible setting. For example, it could be the case that the expression of a gene in cancer blood is greater than in the cancer tissue.*

We aim at discovering genes that exhibit differential expression in all these cases, i.e. we are looking to find the members of the set $\mathcal{C} = C_1 \cap C_2 \cap C_3 \cup C_1 \cap C_2 \cap C_4$. In particular, the genes we consider to be indicative of the alterations in blood for breast cancer patients and that are possibly related to the presence of CTCs should demonstrate up-regulated expression in cancer tissue and cancer blood when compared to normal tissue and blood correspondingly, and also they should have elevated expression in cancerous state both in tissue versus blood and in blood versus tissue.

The purpose and contribution of C_3 is justified based on the consideration of tissue-specific differences in the expression of genes, which have been established in biological studies even for same-condition (homogeneous) population (variability of expression within and across populations) [Kandula et al., 2012, Schobesberger et al., 2008, Radich et al., 2004, Shin et al., 2011]. Basically, the inclusion of C_3 aims to alleviate cross-tissue differences appearing in the base populations of control breast and control blood engaged in our study. Along these lines, if the distributions in these two base populations are similar then C_3 derives similar results as C_1 and cannot contribute any new information. However, if in certain genes there is a large increase either in the first (mean) or second-order (variance) statistic of control blood over control tissue, then the effect is also mapped on the SAM metric used for assessing the differentiation of populations (section 2.2.3). Recall that C_1 compares the distribution of cancer tissue over control tissue, whereas C_3 compares the same distribution over control blood. Thus, even though a gene might present large SAM metric in C_1, this metric can be drastically reduced when the mean or variance of the base population (control blood) increases in C_3, leading to the exclusion of this gene from the C_3 set.

As a final note, we particularly check of "over-expression" only (instead of or in addition to the "under-expression" case) because we want to focus on genes with hyperactivity and elevated expression that are commonly called "oncogenes" [Shastry, 1995]. Oncogenes drive abnormal cell proliferation as a consequence of genetic alterations that either increase gene expression or lead to uncontrolled activity of the oncogene-encoded proteins [Cantley et al., 1991] and have been characterized as the Achilles' heel of cancers [Weinstein, 2002].

Summarizing our motivation, the intersection of sets $C1 \cap C_2$ reveals active genes over the base levels in both tissue and blood, which could be due to cancer causality but also to certain blood differences from tissue. To that respect, the inclusion of C3 relieves the influence of overexpressed genes in blood compared to tissue due to any reasons possibly unrelated to cancer. Such genes are captured in the set $C1 \cap C_2 - C_3$ and excluded from the overall intersection $C_1 \cap C_2 \cap C_3$.

Currently, two prevalent models - progression model and metastatic predestination model - provide evidence about tumor progression towards metastasis [Hunter and Alsarraj, 2009]. Bearing in mind that both models are still under scrutiny and that CTC profiles capture either primary tumor or metastasis molecular characteristics [Barbazán et al., 2012], we applied the two-stage methodological approach described above to derive a panel of genes that are common in primary carcinomas and peripheral blood of breast cancer patients. Overall, we consider the hypothesis that this intersection, representing the common features of primary tumor and breast cancer peripheral blood, is likely to reflect circulating tumor cells biology. In this form, the analysis of the intersection signature might be biologically and therapeutically significant in terms of the involved processes and pathways, forming a useful clinical diagnostic tool.

2.1.2 Related studies

In general, several microarray studies on breast cancer tissue samples (control versus cancer tissue even from the same person) demonstrate alterations in processes manifested in gene deformations. Similar gene alterations appear in the analyzed portion of peripheral blood (control blood versus cancer blood) [Balmain et al., 2003]. In addition, [Barbazán et al., 2012] report that the spread of cancer relates to the detachment of malignant cells into blood and [Obermayr et al., 2010] demonstrate that CTCs can be detected in single-cell level through specific genes (six gene panel) in peripheral blood. Particular microarray studies on peripheral blood that isolate specific CTC cells report that CTCs carry characteristics from the primary cause [Obermayr et al., 2010], but also convey information regarding the secondary (metastasis) tumor [Barbazán et al., 2012]. Thus cancer-specific alterations can be identified in affected tissue areas, as well as in blood. Moreover, some specific alterations in cancer might be indicative of its ability to diffuse; such genes can indirectly predict the existence of CTCs without the need to detect and/or extract them [Molloy et al., 2012].

2.2 Methods and Procedures

2.2.1 Breast Cancer Datasets

We have used 9 different data sets publicly available from the Gene Expression Omnibus (GEO) database [Edgar et al., 2002b], which are shown in Table 2.1 with their relevant characteristics. Most of the data sets provide samples from both normal and cancer breast tissues. Furthermore, there are a variety of different platforms; Affymetrix and Agilent are the most common manufacturers in this collection of datasets while there is one dataset using a custom microarray chip from Agendia and another one from Applied Biosystems (ABI).

2.2.2 Dataset integration

Unfortunately, not all of the three comparisons (C_1, C_2, C_3, and C_4) can be robustly performed by studying a single data set, due to the lack of samples. Even in the case of C_1, which contrasts healthy tissue samples with cancer

Table 2.1: Breast cancer datasets used

GEO Accession	Site	Number of Cancer / healthy samples	Used in
GSE22820 [Liu et al., 2011][a]	Tissue	176 / 10	C_1, C_3, C_4
GSE19783 [Enerly et al., 2011][a]	Tissue	113 / 2 ($0^\dagger$)	C_1, C_3
GSE31364 [Molloy et al., 2012][b]	Tissue	72 / 0	C_1, C_3
GSE9574 [Tripathi et al., 2008][c]	Tissue	14 ($0^\star$) / 15	C_1, C_4
GSE18672 [Haakensen et al., 2010][d]	Tissue	64 / 79	C_1, C_3, C_4
GSE27562 [LaBreche et al., 2011][e]	Blood	57 / 31	C_2, C_3, C_4
GSE16443 [Aaroe et al., 2010][f]	Blood	67 / 54	C_2, C_3, C_4
GSE15852 [Pau Ni et al., 2010][c]	Tissue	43 / 0	C_3
GSE12763 [Hoeflich et al., 2009][d]	Tissue	30 / 0	C_3

[a] Agilent Whole Human Genome Microarray 4x44K G4112F
[b] Agendia human DiscoverPrint v1 custom platform
[c] Affymetrix Human Genome U133A Array
[d] Agilent Whole Human Genome Oligo Microarray G4112A
[e] Affymetrix Human Genome U133 Plus 2.0 Array
[f] ABI Human Genome Survey Microarray Version 2
[†] Two samples in GSE19783 were removed from C1 because they are metastatic
[⋆] 14 samples in GSE9574 are from epithelium adjacent to a breast tumor and were removed

patients, the number of healthy subjects is limited if a single dataset is used. Therefore, we have designed a data integration methodology that combines different datasets in a single multi-platform, multi-origin dataset where the microarray probe intensities have been re-normalized with the removal of study and batch specific variations. This is not a “meta-analysis” based approach because the data are integrated at the gene expression level instead of working on the combination of p-values, effect sizes, and other statistics that have been computed per dataset ([Marot et al., 2009], [Jaffrézic et al., 2007] and [Choi et al., 2003] are examples of such methodologies). Our data integration approach is designated by the heterogeneity in the available datasets in terms of the gene transcripts used and also the biological question and the classification of samples.

The dataset integration approach we have followed consists of the following steps performed for each of the four comparisons:

- Each dataset that is relevant to a given comparison is downloaded from the GEO in the format (e.g. preprocessed) it has been uploaded and registered. However, the raw data are not always available, and it may be the case that some preprocessing tasks have already been performed.

We perform k-nearest neighbors type of imputation [Troyanskaya et al., 2001] if needed and we log-transform the probeset intensities if they have not been already transformed (please refer to Section 1.3 for details).

- For each dataset we map the probes identifiers that are platform specific to Entrez Gene identifiers, in order to have a common "namespace" for the identification of the corresponding genes. Probe sets that lack such a mapping are removed. The mappings used are based on the platform specific annotations in the corresponding Bioconductor libraries [Gentleman et al., 2004]. In the specific case of GSE31364 that uses a custom chip, we have used the annotation provided by GEO using the provided GenBank Accession numbers as the mapping means (Platform GPL14378[1]).

- Since the mapping process explained in the previous step can result in different probes correspond to the same Entrez Gene id, we keep the expression values for the probes that exhibit the largest variation as estimated by the Interquartile Range (IQR). The IQR provides a robust measure of the dispersion of expression values for a given gene, and keeping the probe set with the largest IQR is a kind of univariate "feature selection" to remove non-informative variables.

- After the summarization and probe filter procedure that is performed in the previous step for each study (data set) separately, the method proceeds in each planned comparison by performing a cross-study batch-correction and cross-platform normalization of the comparison relevant datasets using the Empirical Bayes (COMBAT) algorithm [Johnson et al., 2007].

- The final merged dataset for each comparison is subsequently used for finding differential expressed genes by selecting the samples from the original datasets that are relevant to the specific comparison (for example the 31 healthy blood samples of GSE27562 are not used in the C_4 comparison but are used in C_2 and C_3).

2.2.3 Gene differentiation

When the subsets of samples from the relevant data sets have been prepared and merged together, we are using the Significance Analysis for Microarrays (SAM) method [Tusher et al., 2001] with the `siggenes` package of R/Bioconductor in order to discover the genes that exhibit comparison-specific differential

[1] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL14378 (accessed on February, 26 2016)

expression. We are also using the False Discovery Rate (FDR) [Benjamini and Hochberg, 1995b] as the criterion for determining the set of genes that exhibit differential expression and its critical value has been set to 0.01 for all comparisons. The use of FDR implies that the resulting gene sets that were found to have differentiating expression values do not have the same number; instead, the number of differentially expressed genes differs among comparisons.

We are particularly interested in the genes that show significant up-regulation in the various comparisons, so the final step is the selection of those that exhibit one-direction over-expression.

2.3 Results

2.3.1 Statistical comparisons and first findings

Based on the proposed methodology, we have extracted the signatures from the comparison of specific sets of genes, which exhibit differentially over- expressed behavior

What is important at the second step of our process is to find the number and reveal the biological relevance of genes belonging to the different sets, but also in the intersections of these sets. As shown in Table 2.2, for the C_1 comparison, we identify 3,725 genes showing significant differential over-expression in cancer over control tissue. Similarly, for the C_2 comparison, we extract 79 genes over-expressed in cancerous peripheral blood samples. Finally, the C_3 comparison derives 245 genes over-expressed in cancer tissue when compared with control peripheral blood samples whereas C_4 yields 1,076 genes with increased expression in cancer blood versus normal tissue. The size (cardinality) of the different sets of genes alongside with their interestion areas can be seen in the Venn diagram of Figure 2.2. For example, we identify 24 common genes in the $C_1 \cap C_2 \cap C_3$ intersection, 26 in $C_1 \cap C_2 \cap C_4$, 27 genes in the intersection of C_1 and C_2, 137 genes in the intersection of C_1 and C_3, 59 genes in the intersection of C_2 and C_3, etc.

Table 2.2 also shows the genes that have been found in common with other studies, in particular [Molloy et al., 2012] and [Powell et al., 2012]. From Molloy et al. we identify 3 genes (RPS8, WISP1, and TMEM121) in the gene set produced by comparison C_1. Powel et al. report a number of genes that are supportive for the presence of CTCs. Among them we identify Chemokine, CXC motif, receptor 4 (CXCR4) in the common genes of all comparisons, Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) in the intersection of C_1 and C_3, and Vimentin (VIM) in the intersection of C_2 and C_3. Both VIM and CXCR4 have been associated with the epithelial-mesenchymal transition

Table 2.2: Results of the comparisons

Set	Size	Common with [Powell et al., 2012]	Common with [Molloy et al., 2012]
C_1	3725	RT18, KRT19, ACTB, RRM1, S100A9, SLC2A1, TFRC, TGFB1, UBB, CXCR4, CASP3, CD44, CD53	RPS8, TMEM121, WISP1
C_2	79	MAPK14, VIM, CXCR4	
C_3	241	PARP1, MAPK14, GAPDH, VIM, CXCR4	
C_4	1076	PARP1, GAPDH, KRT8, KRT18, ACTB, RRM1, S100A9, SLC2A1, TFRC, TGFB1, UBB, CXCR4, CD44, CD53	
$C_1 \cap C_2$	27	CXCR4	
$C_1 \cap C_3$	134	PARP1, GAPDH, CXCR4	
$C_2 \cap C_3$	59	MAPK14, VIM, CXCR4	
$C_1 \cap C_4$	1035	PARP1, GAPDH, KRT8, KRT18, ACTB, RRM1, S100A9, SLC2A1, TFRC, TGFB1, UBB, CXCR4, CD44, CD53	
$C_2 \cap C_4$	27	CXCR4	
$C_1 \cap C_2 \cap C_3$	24	CXCR4	
$C_1 \cap C_2 \cap C_4$	26	CXCR4	

The intersections of gene signatures with the corresponding number of up-regulated genes. A set of previously identified biomarkers is also mapped into the different sets and intersections.

but there are additional epithelial marker genes like Keratin 8 (KRT8) and Keratin 19 (KRT19) or metastatic genes like calgranulin-B (S100A9) that are found be over expressed in the comparison C_1. In particular S100A9 has been identified as a negative regulator for lymph node metastasis [Choi et al., 2012]. Common to many intersections of the gene sets is the CXCR4 that was recently found to be associated with the mobilization and trafficking of CTCs [Mego et al., 2016].

The exact size and members of the important intersections in the results of the statistical comparisons are shown in Table 2.3.

2.3.2 Biological interpretations

According to our hypothesis, we aim to identify factors in peripheral blood that can indirectly reveal the traffic of circulating tumor cells, instead of CTC detection. This is reflected in the intersection $C_1 \cap C_2 \cap C_3$, in which genes that are over-expressed in normal blood are eliminated and remain only genes that are over-expressed in blood due to cancer-associated factors. This gene signature includes 24 genes as shown in Figure 2.2.

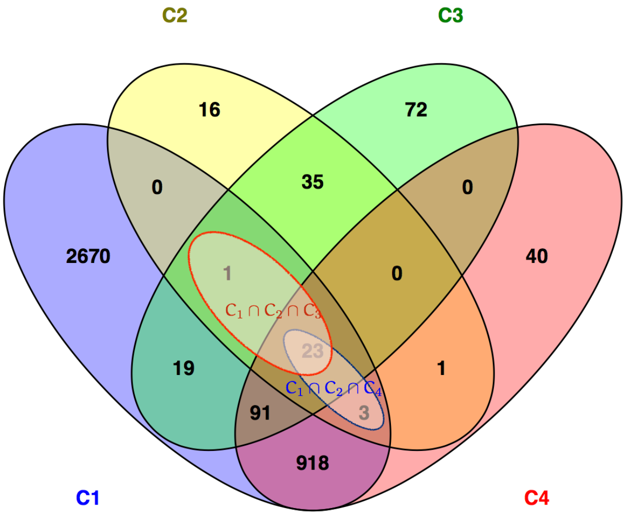


Figure 2.2: The number of differentially expressed genes in the various intersections of the comparisons. As can be seen there are 24 genes in the intersection $C_1 \cap C_2 \cap C_3$*, 26 in the intersection* $C_1 \cap C_2 \cap C_4$*, and 27 genes in the union of these intersections. Finally, there are 23 genes common in all four comparisons. The Venn diagram shown was created through the Venny tool [Oliveros, 2007].*

Table 2.3: Genes identified

Set	Size	Genes
$C_1 \cap C_2 \cap C_3$	24	TRIB1, CDKN2D, TMED10, GABPB1, GALK2, GLO1, HMGN2, EIF6, JUNB, KPNA4, NFYA, PRDX1, WDR83OS, TMEM70, WIPI1, SAR1A, SRSF6, TYROBP, YWHAB, CXCR4, DHX58, BECN1, MAFB, PTBP3
$C_1 \cap C_2 \cap C_4$	26	TRIB1, CDKN2D, TMED10, GABPB1, GALK2, GLO1, HMGN2, HNRNPU, JUNB, KPNA4, NFYA, PRDX1, WDR83OS, TMEM70, WIPI1, PRKAR1A, SAR1A, SRSF6, SNRPF, TYROBP, YWHAB, CXCR4, DHX58, BECN1, MAFB, PTBP3
$C_1 \cap C_2 \cap C_3 \cap C_4$	23	TRIB1, CDKN2D, TMED10, GABPB1, GALK2, GLO1, HMGN2, JUNB, KPNA4, NFYA, PRDX1, WDR83OS, TMEM70, WIPI1, SAR1A, SRSF6, TYROBP, YWHAB, CXCR4, DHX58, BECN1, MAFB, PTBP3
$C_1 \cap C_2 \cap C_3 \cup C_1 \cap C_2 \cap C_4$	27	TRIB1, CDKN2D, TMED10, GABPB1, GALK2, GLO1, HMGN2, HNRNPU, EIF6, JUNB, KPNA4, NFYA, PRDX1, WDR83OS, TMEM70, WIPI1, PRKAR1A, SAR1A, SRSF6, SNRPF, TYROBP, YWHAB, CXCR4, DHX58, BECN1, MAFB, PTBP3

Classification of all four gene signatures ($C_1 \cap C_2$, $C_2 \cap C_3$, $C_1 \cap C_3$, and $C_1 \cap C_2 \cap C_3$) was conducted by WebGestalt (WEB-based GEne SeT AnaLysis Toolkit) [Zhang et al., 2005] in order to evaluate the most enriched Kyoto

Encyclopedia of Genes and Genomes (KEGG) pathways and gene-ontology (GO) terms for the category of biological processes (BP). WebGestalt utilizes the hypergeometric test for the enrichment of GO and KEGG terms in the selected genes, followed by the Benjamini & Hochberg (BH) method for multiple test adjustment (adjP) [Benjamini and Hochberg, 1995a]. Additionally, we used the G2SBC (Genes-to-Systems Breast Cancer) Database [Mosca et al., 2010b], which is a valuable resource that integrates gene-transcript-protein data reported in literature as altered in breast cancer cells, to annotate all genes that are associated with breast cancer in each of these gene signatures.

The biological results correlate well with the statistical results. The intersections $C_1 \cap C_2$, $C_2 \cap C_3$, and $C_1 \cap C_3$ derive gene signatures with 3, 35 and 113 genes, respectively, without taking into account the all-common genes of $C_1 \cap C_2 \cap C_3$. Importantly, all these signatures include breast cancer associated genes in a relative high percent ($C_1 \cap C_2 = 33.3\%$, $C_2 \cap C_3 = 20\%$, and $C_1 \cap C_3 = 33.63\%$) according to G2SBC database. Moreover, according to WebGestalt the enriched biological processes include: for $C_1 \cap C_2$, RNA processing (adjP< 0.05); for $C_2 \cap C_3$, biosynthetic process, protein modification process, metabolic process, gene expression, Toll signaling pathway and osteoclast differentiation (adjP< 0.05); and for $C_1 \cap C_3$, antigen processing and presentation, response to stress and cell cycle (adjP< 0.05). In addition, the most enriched KEGG pathways include: for $C_1 \cap C_2$, spliceosome (adjP=2.58e-05); for $C_2 \cap C_3$, Toll-like receptor signaling pathway (adjP=1.90e-05) and osteoclast differentiation (adjP=2.89e-05); and for $C_1 \cap C_3$, proteasome (adjP=4.18e-06) and lysosome (adjP=1.57e-05).

Finally, focusing on the 24 genes of $C_1 \cap C_2 \cap C_3$ intersection, we observe that seven of them are associated with breast cancer and the most enriched biological processes (adjP < 0.05) include RNA splicing, and the autophagic vacuole assembly (adjP $= 0.0432$). In addition, the majority of genes participate in regulation of the metabolic process; however, without reaching statistical significance. The cell cycle and osteoclast differentiation are both enriched (adjP $= 0.0023$) KEGG pathways in $C_1 \cap C_2 \cap C_3$ intersection. The biological enrichment associations of the 24 genes of this intersection are presented in Table 2.4. In addition, the biological enrichment associations of the 27 genes of the $C_1 \cap C_2 \cap C_3 \cup C_1 \cap C_2 \cap C_4$ set are presented in Table 2.5.

Desmedt et al [Desmedt et al., 2008] note that different prognostic signatures are not evaluated and compared on similar molecularly defined subgroups (e.g. ER and HER2 subgroups) although the original studies address the same questions, so that there is no overlap or only little between their gene lists. Targeted studies have specified major biological processes in breast cancer, such as proliferation, tumor invasion/metastasis, impairment of immune response, evasion of apoptosis, self-sufficiency in growth signals, and ER/HER2

Table 2.4: Gene Ontology (GO) Biological Processes

Biological Process	Adjusted P-value (BH)	Affected Genes
negative regulation of RNA splicing [GO:0033119]	0.0378	PTBP3, SRSF6
organelle assembly [GO:0070925]	0.0378	EIF6, WIPI1, BECN1
peptide metabolic process [GO:0006518]	0.0378	GLO1, TMED10, BECN1
erythrocyte homeostasis [GO:0034101]	0.0378	PTBP3, MAFB, PRDX1
beta-amyloid metabolic process [GO:0050435]	0.0378	TMED10, BECN1
autophagic vacuole assembly [GO:0000045]	0.0432	WIPI1, BECN1
vesicle targeting, to, from or within Golgi [GO:0048199]	0.0432	WIPI1, TMED10
cellular amide metabolic process [GO:0043603]	0.0529	GLO1, TMED10, BECN1
regulation of cellular metabolic process [GO:0031323]	0.0529	NFYA, CXCR4, WIPI1, SRSF6, JUNB, PTBP3, MAFB, GLO1, CDKN2D, BECN1, GABPB1, YWHAB, TRIB1, HMGN2
macroautophagy [GO:0016236]	0.0529	WIPI1, BECN1

signaling [Hanahan and Weinberg, 2000]; but other key biological processes are likely to be added to this list in the future. Indeed, our 24 gene signature consists of both known and emerging features of cancer, namely the autophagy and the reprogramming of energy metabolism. Autophagy, which is involved in our signature as elements of the autophagic program (e.g. biosynthesis, energy metabolism, intracellular vesicles, lysosomes), represents an important cell-physiologic response and is known to mediate both tumor survival and death [Hanahan and Weinberg, 2011], while the reprogramming of energy metabolism, which is represented in Table 2.4 by cellular amide metabolic process and peptide metabolic process, was added as one of two emerging hallmarks of potential generality to the cancer list. In addition, a recent study in [Lozy and Karantza, 2012] emphasizes the continuous interplay of reprogrammed cancer cell metabolism and autophagy, which is modulated by many tumor related conditions including oxidative stress (beta-amyloid possibly increase the generation of reactive oxygen species, GO: 0050435, Table 2.4). The latter allows cancer cells for rapid adaptation to stressful environmental conditions, preservation of the uncontrolled proliferation, as well as prevention of toxic radiation and/or chemotherapy effects [Lozy and Karantza, 2012].

Moreover, Shi et al [Shi et al., 2010] performing a co-expression module

Table 2.5: Gene Ontology (GO) Biological Processes

Biological Process	Adjusted P-value (BH)	Affected Genes
peptide metabolic process	0.0278	GLO1, TMED10, BECN1
autophagic vacuole assembly	0.0278	WIPI1, BECN1
positive regulation of cellular metabolic process	0.0278	CXCR4, NFYA, WIPI1, JUNB, MAFB, BECN1, GABPB1, YWHAB, PRKAR1A, TRIB1
ribonucleoprotein complex assembly	0.0278	SNRPF, EIF6, SRSF6
erythrocyte homeostasis	0.0278	PTBP3, MAFB, PRDX1
negative regulation of RNA splicing	0.0278	PTBP3, SRSF6
organelle assembly	0.0278	EIF6, WIPI1, BECN1
vesicle targeting, to, from or within Golgi	0.0278	WIPI1, TMED10
beta-amyloid metabolic process	0.0278	TMED10, BECN1
cellular nitrogen compound metabolic process	0.0351	SRSF6, SNRPF, MAFB, GLO1, BECN1, GABPB1, TRIB1, HMGN2, NFYA, SAR1A, TMED10, WIPI1, JUNB, PTBP3, HNRNPU, CDKN2D, YWHAB, PRKAR1A
intracellular transport	0.0401	KPNA4, WIPI1, TMED10, SAR1A, SRSF6, YWHAB, PRDX1
cellular amide metabolic process	0.0401	GLO1, TMED10, BECN1
regulation of cellular metabolic process	0.0401	CXCR4, NFYA, WIPI1, SRSF6, JUNB, PTBP3, MAFB, GLO1, CDKN2D, BECN1, GABPB1, YWHAB, PRKAR1A, TRIB1, HMGN2
macroautophagy	0.0401	WIPI1, BECN1
transcription from RNA polymerase II promoter	0.0401	MAFB, SNRPF, GLO1, NFYA, GABPB1, SRSF6, JUNB, PRKAR1A
termination of RNA polymerase II transcription	0.0401	SNRPF, SRSF6
RNA splicing	0.0401	PTBP3, SNRPF, HNRNPU, SRSF6
cellular macromolecular complex assembly	0.0401	SNRPF, EIF6, TMEM70, TMED10, SRSF6
regulation of protein phosphorylation	0.0401	CXCR4, CDKN2D, WIPI1, YWHAB, PRKAR1A, TRIB1
homeostasis of number of cells	0.0413	PTBP3, MAFB, PRDX1
mRNA metabolic process	0.0468	PTBP3, SNRPF, HNRNPU, SRSF6, YWHAB
establishment of vesicle localization	0.0468	WIPI1, TMED10
Golgi vesicle transport	0.0468	WIPI1, TMED10, SAR1A

analysis reveal biological processes that are associated with breast cancer progression. They found three groups of modules, one of which (Group

II) included up-regulated modules such as cell cycle, RNA splicing, cellular component organization and protein metabolic process that are related to uncontrolled cell proliferation, a hallmark of cancer. All these processes have been found in our $C_1 \cap C_2 \cap C_3$ intersection forming the CTC-related 24 gene signature.

Nowadays it is also known that the mechanisms of cell-cycle, a pathway that we found in our $C_1 \cap C_3$ and $C_1 \cap C_2 \cap C_3$ intersection signatures, are deregulated at multiple levels in breast cancer cells [Caldon et al., 2006]. Finally, the KEGG pathway of osteoclast differentiation was also found in our 24-gene signature through TYROBP, and JUNB (Table 2.4). This process has been found to be stimulated by a novel factor (CCN3), which impairs osteoblast differentiation to promote breast cancer metastasis to bone [Ouellet et al., 2011]. Thus it is interesting to further consider the association of genes such as TYROBP and JUNB, with breast cancer metastasis. Notice that, DAP12 (TYROBP) is substantial for macrophage fusion, and the production and function of osteoclasts, while its breast cancer expression is more recently connected with bone and liver metastases [Shabo et al., 2013].

In a related framework Hanahan and Weinberg communicate specific biological capabilities that constitute the hallmarks of cancer and are acquired during multistage tumor development in humans, which include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, reprogramming of energy metabolism and evading immune destruction [Hanahan and Weinberg, 2000,Hanahan and Weinberg, 2011]. These distinctive attributes form a structured principle to streamline the complex nature of neoplastic diseases. They can be explored through the observation of a set of perturbed genes in microarray experiments, but they should be confirmed in engaged biological pathways and processes or correlated to risk categories [Wirapati et al., 2008, Fan et al., 2006, Molloy et al., 2012] and through such enrichment analysis to assess their relevance to tumor development, progression, invasion and metastasis [Hung, 2013].

Based on the above observations, we conclude that all four gene signatures contain valuable information regarding breast cancer disease. Specifically, the 24-gene signature, which is expected to involve all these factors that are associated indirectly with the circulation of tumor cells, appears to fulfill our exploration. In our study we demonstrated that our two-step process provides a 24 gene signature of the $C_1 \cap C_2 \cap C_3$ intersection with main components of breast cancer characteriscs and with good association to the characteristics of CTCs.

In summary, we attempted to identify a common list of genes in primary tumor tissues and breast cancer peripheral blood from breast cancer patients.

Using gene set enrichment analysis we identified key pathways and biological processes that are well known to be implicated in breast cancer and metastasis; i.e. they have a biological association to deregulated mechanisms of breast cancer and can possibly reflect the CTC status.

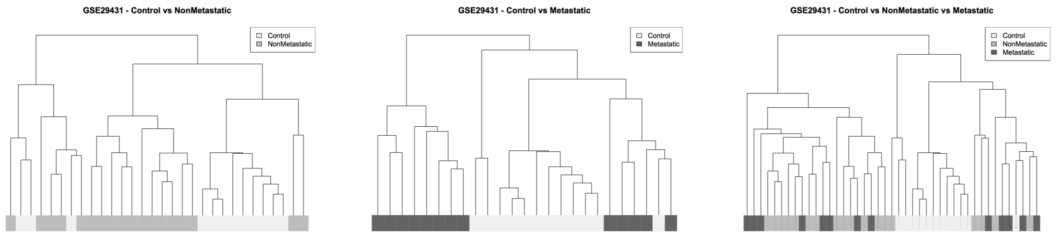
2.4 Independent Validation of Results

2.4.1 Classification performance in Independent Datasets

Recent studies highlight the significant role of the external validation of signatures that have been generated from specific datasets, as to allow repeatability on other datasets [Kim and Kim, 2008]. In order to gain further insight on our results, we test the performance of the 24 genes in the intersection of C_1, C_2, and C_3 on two independent datasets, (GSE29431 and GSE42568), which of cource have not been used for the elicitation of our gene sets. We also evaluate these genes on the already used dataset (GSE19783) but on different class labelling, focusing on the discrimination of circulating and disseminating tumor cells.

The first independent dataset (GSE29431) by Lopez et al. [Lopez et al., 2012] provides microarray data from 54 primary breast carcinomas and 12 samples of breast normal tissues from breast cancer patients. Considering the information about metastatic status we only include 31 tumor samples (18 metastatic, 13 non- metastatic) and all 12 samples of breast normal tissues for validation. The second study by Clarke et al. [Lopez et al., 2012] provides microarray data (GSE42568) from 104 cases of primary breast cancer, prior to any treatment with tamoxifen or chemotherapeutic agents, with known metastatic status of lymph nodes (forty- five tumors with no axillary metastases and 59 tumors showing metastasis to axillary lymph nodes), and 17 samples of normal breast tissues. Both datasets (GSE29431 and GSE42568) are available from GEO [Barrett et al., 2005] and measured on the same platform (Affymetrix Human Genome U133 Plus 2.0 Array).

In order to validate our methodology, we use hierarchical clustering on the first dataset (GSE29431) and we show that the 24 genes that we identified can effectively separate the population of control (healthy) from tumor samples. We consider the differentiation between control and non-metastatic, as well as control and metastatic populations. In order to assess the grouping of samples, we also present the correct class label encoded in black and gray color in a separate row. The distribution of labeled samples in clusters is graphically depicted in Figure 2.3a and 2.3b, for the comparisons of control with non-metastatic and metastatic populations, respectively. The control population shows different characteristics that enable the inclusion of most samples (9 from 12 in each test) in a single cluster. Furthermore, the hierarchical clustering of all three populations is presented in Figure 2.3c, where we observe only partial separation of the two cancer classes, as expected from their biological context [Hunter and Alsarraj, 2009, Weigelt et al., 2003, Choi et al., 2012]. There are contradictory findings regarding the molecular characteristics of the primary tumors and their metastases, but it is widely assumed that the gene expression profiles of metastases are broadly similar to that of primary breast carcinomas [Hunter and Alsarraj, 2009, Weigelt et al., 2003, Wu et al., 2008].



(a) Control vs Non-Metastatic *(b) Control vs Metastatic* *(c) Control vs Metastatic and Non-Metastatic*

Figure 2.3: The results of the hierarchical clustering for the dataset GSE29431.

To elaborate more on these issues, we also examine the second dataset (GSE42568). For the larger number of cases in this dataset, we test the predictive ability of the 24 genes of interest by examining the classification accuracy in a leave-one-out cross validation (LOOCV) scheme. We test regression approaches including “Least Absolute Shrinkage and Selection Operator” (LASSO) and “Support Vector Machines” (SVM) as classifiers applied on pairs of populations. The LASSO classifier achieves 95,16% LOOCV accuracy for the pair-test of (control vs non-metastatic) and 96,05% for the test of (control vs metastatic). The classification accuracy drops to 63,46% on the cancer test of populations (metastatic vs non-metastatic). The SVM approach achieves lower rates but with the same order corresponding to 93,55%, 89,47% and 57,69% for the above pair comparisons, respectively. As in the previous case, the 24 gene signature enables the discrimination of control from cancer populations on independent datasets, but they enable only partial discrimination of metastatic from non-metastatic cases.

As a last effort in searching for the “most informative” genes that can possibly characterize metastasis, we examine the LASSO approach within a recursive feature elimination strategy, where the less significant feature is eliminated at each iteration. Note that each iteration achieves a full test of LOOCV with the remaining features (genes). The maximum accuracy (65,38%) in the test including the metastatic vs non-metastatic populations is achieved with 11 genes in the list (TRIB1, GLO1, HMGN2, EIF6, JUNB, WIPI1, CXCR4, DHX58, BECN1, MAFB and PTBP3). Since the ultimate goal of our efforts is to assess the role of blood markers in metastasis and especially in CTCs, we further examine the role of the 24 genes in disseminating and/or circulating tumor cells associated with cancer. More specifically, we consider the gene expression data of primary tumor in the dataset (GSE19783) (based on information available from previous studied Dataset GSE3985) and use them to predict the DTC status in a LOOCV classification environment. Starting from the proposed list of 24 genes, we apply recursive feature elimination and compare the predicted with the verified status from actual bone marrow aspirates of the same patients. Interestingly, the maximum accuracy (73,8%) is achieved for 10 genes, five of which are common with the previous 11 genes characterizing metastatic from non-metastatic samples. These genes include (TRIB1, GLO1, WIPI1, CXCR4 and BECN1). We perform the same analysis for the CTC labels of this dataset, as presented in [Molloy et al., 2012]. Notice that the analyzed samples are from breast cancer tissue, whereas the CTC labels are determined using multi-marker QPCR-based CTC assays for peripheral blood samples from the same subjects [Molloy et al., 2012]. The RFE-LASSO approach in LOOCV scheme achieves 77,8% accuracy with 12 genes. This list includes 5 common genes (TRIB1, CDKN2D, WIPI1, YWHAB and CXCR4 genes) with the previous comparison of DTCs showing good consistency with the CTC status. Overall, these independent tests may assess the importance of the extracted panel of 24 genes not only in cancer but also in metastasis through disseminating or circulating tumor cells. This is also supported by the fact that the reported genes have been detected and extracted as commonly overexpressed in cancer for both breast tissue and blood, emphasizing the association of peripheral blood in metastasis and indicating good prospects in detecting CTCs and cancer-causal molecules through the profiling of bulk blood samples.

2.5 Discussion

The pairwise signatures and their integration examined in this study derive genes involved in significant cancer processes related to aggressiveness and metastatic behavior. As such, they can be further studied for the assessment of the presence of CTCs in peripheral blood, without the need of isolating and processing single cells. Finally, the comparisons we perform for the derivation of signatures C_1, C_2, C_3, and C_4 are more or less similar to those in many other studies. The interesting aspect of our study is the integration of many datasets, especially for increasing the size of the control group. A similar intersection of the form of $C_1 \cap C_2$, has been also implemented in [Obermayr et al., 2010]. The addition of the C_3 in the intersection process is has been found essential in order to exclude irrelevant genes that are expressed in cancer peripheral blood but are not specifically related to cancer.

The data integration methodology we illustrate above can be criticized for various reasons. First, the merger and integration of platform heterogeneous data sets endangers the elimination of probesets that do not map into the common set of gene identifiers for all the data sets of a given comparison. For example, the Affymetrix U133 Plus 2.0 Array supports up to 20,000 unique gene identifiers but the ABI chip maps only to around 7,000 genes. Thus, the final combined dataset in C_2 can contain up to the smaller number of genes. In fact, the final dataset can possibly contain a lot less measured gene transcripts because there can be an even smaller set of common Entrez Gene identifiers. Since there is no real remedy for this problem, we use as similar and complete platforms as possible in every comparison we make.

Secondly, the cross-study normalization of the expression values can be too pervasive, making the merged data sets overly similar by eliminating potentially important variations. An example of the effect of such (re)normalization across datasets in the same comparison can be seen in Figure 2.4, where the healthy samples from one study are shown before and after the merger. The expression values of different genes appear to be shifted so that their central tendency statistics (mean, median) come closer but the gene-wise variability is almost retained. Such a transformation in the expressions of genes is generally preferred when compared to no additional normalization: in the latter case the unwanted study-specific variability, which is usually systematic (i.e. "batch effects"), can be overwhelming [Leek et al., 2010]. As explained in the Methods paragraph, a "meta-analytic" approach is not always possible although it can be more robust since the platform and study-specific differences become irrelevant by avoiding the integration of raw expression values. We have therefore adopted the integration of the studies at the expression level using the COMBAT algorithm, which appears to yield the best results in various cases and retains legitimate biological variation between the biologically distinct samples [Tsiliki et al., 2011, Turnbull et al., 2012].

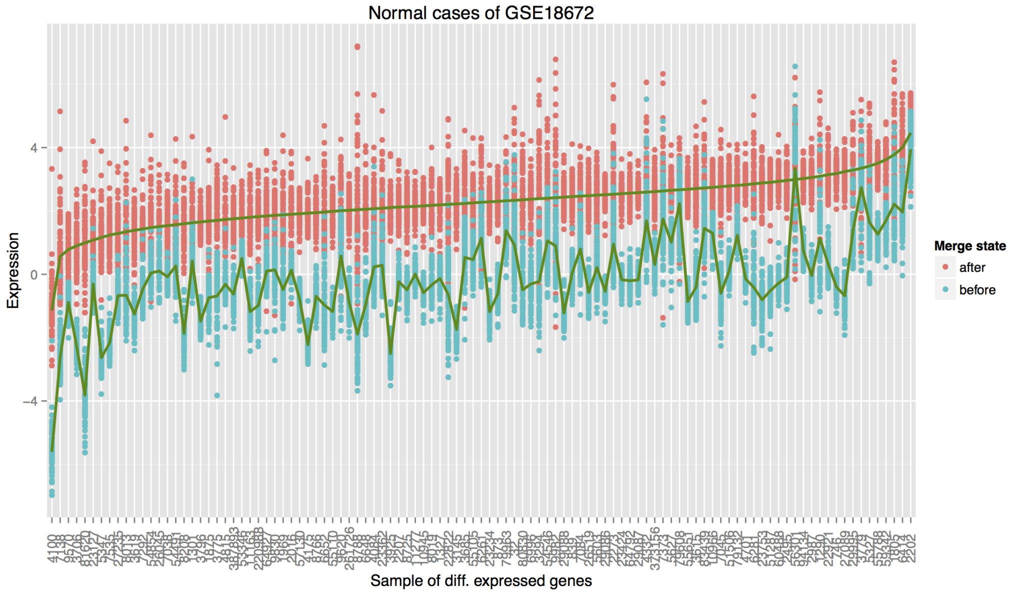


Figure 2.4: An example of the effect of the cross-study normalization. The expression of a random sample of the differentially expressed genes in the C_1 *comparison is shown before and after the Empirical Bayes (COMBAT) method is performed in the normal samples of the GSE18672. The gene-wise variation seems to be preserved but the gene-wise mean expression value (shown as a green line) is smoother after the merger.*

2.6 Conclusions

The identification and characterization of CTCs is important for the treatment of patients with metastatic epithelial cancers such as breast cancer. The isolation of CTCs though is difficult due to the small numbers of such disseminating cells in the peripheral blood. In this chapter we describe a multi-study integration approach that attempts to explore the field by combining microarray gene expression data originated from tissue and peripheral blood.

The results are indeed promising. The 27-gene signature contains some of the important genes that are commonly used for CTC identification, and therefore it makes sense to further consider the proposed approach for indirect assessment of the existence of CTCs.The Supplementary Material and the R/Bioconductor code for the present analysis are hosted in GitHub and can be found at `https://github.com/sgsfak/data_int_ctc`.

In the subsequent chapters we expand this list of potential biomarkers using biological networks. The objective is to strengthen the discriminating and predictive power of these biomarkers and to get a better underdstanding of the underlying biological mechanisms that give rise to their prominence in the analysis we performed.

Chapter 3

Network analysis

Contents

The result of our work so far is a list of 27 genes that are potential biomarkers for the existence of circulating tumor cells in the blood of breast cancer patients. These genes were found as the outcome of a statistical analysis of public data sets, where (hidden) confounding factors or "noisy" observations may exist, and therefore are introduced with some level of uncertainty. In order to validate and extend these findings, in this chapter we incorporate existing biological knowledge in terms of biological networks. Our first aim is to investigate whether the 27 genes can be used as a module of biological functionality, that is to say, if there are "connected", since there is evidence that disease genes tend to cluster together and co-occur in central network locations [Ideker and Sharan, 2008]. As a second objective, we use the list of the 27 genes as "landmarks" or "seeds" in order to enrich this set of biomarkers and increase their predictive and discrimination ability.

3.1 Introduction

In this chapter, we develop a stepwise refinement approach for biomolecular-network construction utilizing 27 genes characterized as "putative markers of circulating tumor cells" (CTCs) in order to provide the molecular information on breast cancer origin, and its potential to spread to other areas of the body, such as to the brain. A key element in our approach is the consideration of interaction network effects of specific genes suspect for revealing the presence of circulating tumor cells in peripheral blood that are considered the primary cause for the dissemination of cancerous cells [Greene et al., 2012, Dong et al., 2013]. In the previous chapter we have singled out a number of genes that appear to be correlated with the presence of CTCs in breast cancer patients. Here, we use these "biomarkers" as the input alongside with the biological network information in order to better understand the mechanisms that may have given rise to their manifestation. The objective is twofold: a) to gain insight for the biological underpinnings that can possibly give rise to the set of biomarkers and explain their prominence, and b) to expand these biomarkers to a larger pool of genes in a computational, data driven approach and explore the prognostic ability of the derived gene signature in similar but independent datasets.

We are using the biological networks of proteins (Protein-Protein Interactions, Section 1.5) as the source of well-established biological knowledge. The interaction networks provide a lot more potential and flexibility for uncovering hidden associations among the genes than other sources of biological information such as the Gene Ontology, due to their graph-based structure. Furthermore, they have been the subject of a lot of contemporary and important research. Usually, proteins (and therefore the genes that encode them) associated with the similar phenotypes (e.g. disease) are highly interconnected [Jonsson and Bates, 2006]. Also, disease genes have been reported to form strongly connected modules in "central" locations in the networks [Ideker and Sharan, 2008] or to encode proteins with a lot of connections ("hubs") [Wachi et al., 2005]. Due to this extended evidence, in this chapter we consider the list of the 27 genes as "seeds" for the construction of a single module of related biological functionalities, which can possibly include many other interconnected genes.

3.2 Materials and Methods

3.2.1 Methodology

A list of 27 genes derived from our earlier work [Sfakianakis et al., 2014] is our starting molecular signature. These genes were the outcome of the multiple comparisons between normal and cancerous samples in blood and breast tissues. We refer to this set of genes as the input or “seed” list from now on.

In order to understand better the possible biological mechanisms behind the selection of the specific genes in the seed list the first step in our methodology is to construct a network that spans these genes. This network construction step uses known gene (protein) interactions in order to interconnect the genes by introducing additional intermediate vertices in a network. This is known as the Steiner tree problem in graphs [Winter, 1987].

The next step is to expand the elicited network in a data driven way. We use the background biological network that was also employed in the first step but this time combined with a gene expression (GE) dataset in order to introduce additional nodes on the periphery of the Steiner tree. The selection of the neighboring genes to attach to the current network is based on their ability to increase the association of the currently expanding subnetwork with the class labels of the GE dataset’s samples. This step results in a bigger network that we subsequently test its discrimination power in an independent dataset.

3.2.2 Network construction

The Steiner tree for an undirected distance graph $\mathsf{G} = (\mathsf{V}, \mathsf{E}, \mathsf{d})$ and a subset of vertices $\mathsf{S} \subseteq \mathsf{V}$ (called Steiner points from now on) is a connected tree T, with vertices $\mathsf{U} \subseteq \mathsf{V}$ and edges $\mathsf{S} \subseteq \mathsf{E}$ that spans all vertices in S. A minimal Steiner tree for G and S is the one with the minimal total distance of its edges among all the similar Steiner trees. Therefore, the Steiner tree problem aims to find a minimum cost solution for connecting a subset of the graph’s nodes through the selection of some of the “internal” nodes of graph, whereas in the (similar) minimum spanning tree problem the objective is the selection of edges to minimally interconnect all the nodes of the graph. Finding a minimal Steiner tree is an NP-Complete problem and in this study we are using the heuristic algorithm of Kou et al. [Kou et al., 1981] that has been shown to produce trees that are not far from the minimal (optimal) solution.

As the initial Steiner points we use the set of genes in our seed list. We also use the HINT database that provides high-quality protein-protein interaction networks in human and other organisms [Das and Yu, 2012]. Twenty-three (23) of the genes in the seed list were found in the HINT human PPI network and the Steiner tree finding algorithm produces a 56-gene network shown in Figure 3.1. As can be seen, the result is an acyclic graph (a tree) that spans and connects the genes in the seed list by introducing intermediate nodes according to the underlying graph in the HINT database

Table 3.1: Databases and Data sets

Database	Number of nodes	Number of Edges
HINT	10889	45226

Dataset	Platform	Samples
GSE42568	Affymetrix Human Genome U133 Plus 2.0 Array	17 normal and 53 grade 3 breast cancer patients
GSE52604	Agilent-014850 Whole Human Genome Microarray 4x44K	35 Breast Brain Metastasis samples, 10 Non- Neoplastic Brain samples, and 10 Non-Neoplastic Breast samples

3.2.3 Network expansion

The algorithm of Chuang [Chuang et al., 2007] is subsequently used to expand the Steiner tree network and explore the neighborhood of the Steiner tree nodes. To this end we are using the public dataset of GSE42568 (Table 3.1) and the subset of "Grade 3" breast cancer samples versus the normal cases, as follows:

- Each of the genes in the induced Steiner tree is used as a "seed" for possible network expansion.
- For each "seed" the methodology in Chuang et al. is followed: In an iterative fashion, each of the seed's neighbors in the HINT PPI network is added to the subnetwork rooted at the seed and the set of subnetwork nodes is checked whether the agreement with the class labels (i.e. breast cancer versus normal cases) is improved. To determine the agreement with the samples' phenotypes, an "activity score" for the current subnetwork is computed based on z-transformed scores of the genes and the Mutual Information (MI) between the activity score and the class labels is computed. The subnetwork for a specific seed is expanded by checking its neighbors, their neighbors, and so on as long as the MI score is continuously increased.
- The expanded subnetworks from the genes of the Steiner tree are then "sewn" together to form a bigger, unified network that has the original Steiner tree as its "backbone".

Using the Steiner tree of Figure 3.1, the HINT PPI underlying network,

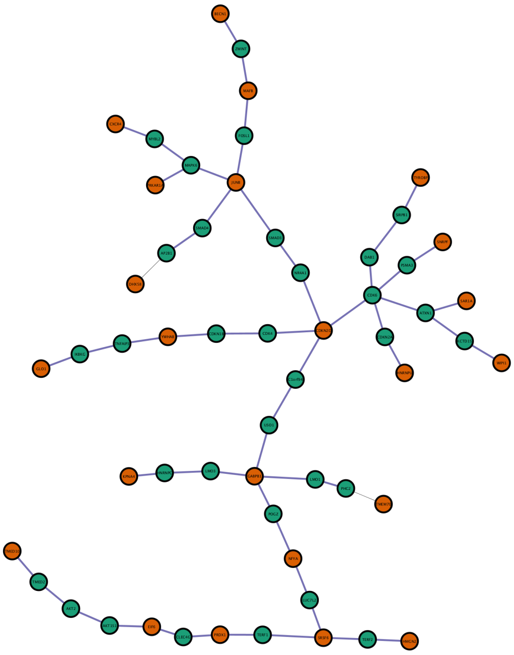


Figure 3.1: The Steiner tree of the input "seed list". The genes used as the initial Steiner points (the seed list) are shown with red fill color, whereas with green color are the additional connecting genes.

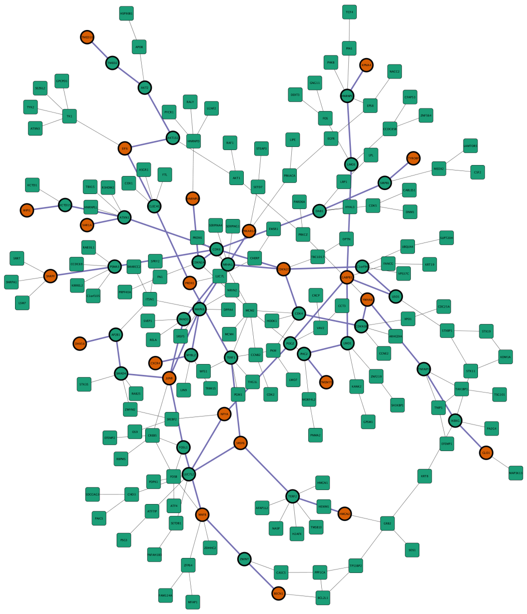


Figure 3.2: The expanded network based on the Steiner tree of Figure 3.1. The genes contained in the original Steiner tree are shown with circles.

and the samples of GSE42568 described in Table 3.1, the above methodology produces a 203-gene network that is shown in Figure 3.2.

3.2.4 Validation of the induced network

We explore the discrimination abilities of the produced network in the independent dataset [Salhia et al., 2014] available from GEO under the accession number GSE52604 (Table 3.1). This dataset contains three types of samples: non-neoplastic brain tissues, non-neoplastic breast tissues, and breast brain metastatic samples. We are interested in exploring the discriminating power of the produced network in two comparisons: a) between the non-neoplastic breast tissues and the breast brain metastatic ones, and b) between the non-neoplastic brain tissues and the breast brain metastases.

A "heatmap" visualization of the expression values for the genes of out network clearly shows that there are differences between these types of samples (Figure 3.3). We subsequently perform gene-wise t-tests for our network in the two comparisons with multiple test adjustment based on the False Discovery Rate (FDR) [Benjamini and Hochberg, 1995a]. With the FDR criterion set at 0.1 the comparison between the non-neoplastic breast and the breast brain metastases yield 52 differential expression genes. Of these, 10 genes are in the identified Steiner tree while 6 genes come from the original seed list we used as input. The similar comparison between the non-neoplastic brain and the breast brain metastatic tissues results in 74 differentially expressed genes, 19 of which are nodes in the Steiner tree and 11 genes belong also to the input gene list.

We next perform a stratified 10-fold cross validation (CV) and aggregate the classification accuracy of the penalized generalized linear model using as features only the 203 genes of the induced network. We are using the glmnet package in the R statistical software for fitting the logistic regression models with the "elastic-net" penalty [Friedman et al., 2010]. The comparison between the non-neoplastic brain and the breast brain metastases returns a mean (along the 10 folds) classification accuracy of 0.98 whereas the corresponding for the breast cases yields the optimum 1.0. To further validate the statistical accuracy of these classification results we perform a simple permutation test [Ojala and Garriga, 2009]: 1000 hypothetical datasets are created by permuting the class labels of the samples. For each of these perturbed datasets we evaluate the classification power of the same genes with the same 10-fold CV methodology. The results can be seen in Figure 3.4 alongside with the real classification accuracy in the original dataset. Under the null distribution induced by these permutation results, the empirical p-value of the real classification results is around 0.001 and therefore the reported classification performances are

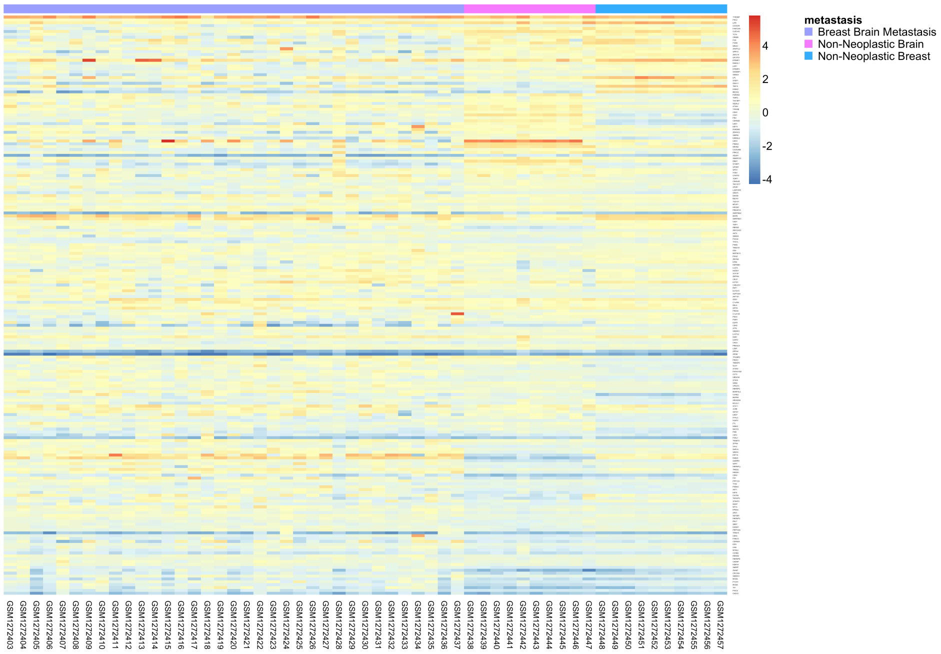


Figure 3.3: The heatmap of the 203 genes of the expanded network in the GSE52604 dataset.

statistically significant.

3.3 Discussion

The proposed method uses network information twice: once for constructing a network connecting the given set of biomarkers and then for expanding the constructed network. The first step is network driven only and therefore common for all cases (e.g. different physiologies, diseases, etc). In the second step though we also take into account specific gene expression data from patients' samples and therefore is more data-driven and case or disease specific. Therefore the method combines some static knowledge in the form of biological interaction networks and dynamic knowledge that is patient and disease specific. We argue that this is an important feature of the proposed methodology

During the first step of our method we try to identify network paths that can interconnect the input list of genes. To this end the Steiner tree algorithms try to solve the optimization problem of finding a small number of intermediate genes. Of course there might be multiple alternative trees connecting the genes, so why opt for the minimal (or near minimal) number of interconnecting genes? We could indeed consider all possible ways that the

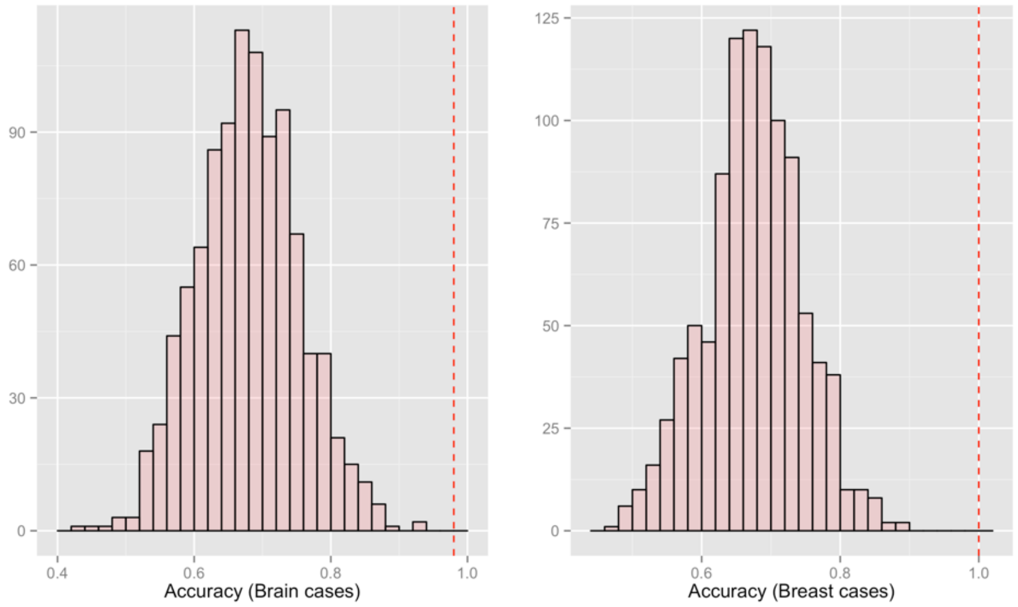


Figure 3.4: Accuracy results when performing 1000 permutations of class labels. The actual performance of the 203 genes is shown with dashed, red vertical lines.

input list of genes connect to each other but this can increase exponentially the "solutions" to be tested. Irrespective of these computational issues, the formulation of the gene interconnection as a Steiner tree problem offers the advantage of the introduction of a little number of additional nodes. This is in agreement with the Occam's razor in the sense that we are searching for the simpler justification on how the input genes have been selected in the first place: the simpler, shorter, and with the fewest assumptions explanation is usually the correct one.

From a biological point of view, following this first step, and analyzing the gene sets by Genes-to-Systems Breast Cancer to identify disease pathways, we observed that the intermediate vertices (AKT2 CDK4 CDK6 CDKN1B CDKN2A) of the "backbone" Steiner tree possess a central role in the constructed protein network by interconnecting the 27 genes of the starting molecular signature and thus supporting the notion that the encoded proteins may be important for network's integrity. The constructed "backbone" Steiner tree unfolds a panel of various significant ($0.000001809 \leqslant p \leqslant 0.008864$) cancer disease pathways that are related to various forms of cancer (Chronic myeloid leukemia, Pancreatic cancer, Small cell lung cancer, Non-small cell lung cancer, Melanoma, Glioma) apart from other important and significant ($0.00009544 \leqslant p \leqslant 0.0452$) molecular pathways (p53 signaling pathway, Signaling by TGF beta, Cell cycle) that are important in health and disease.

For the network expansion step we are using the greedy search based on the Mutual Information that was initially proposed by Chuang et al. Alternative approaches exist, for example Yang et al. propose the "EgoNet" [Yang et al., 2014]. Their technique involves the expansion of the current sub-network by incorporating all the neighboring genes and then testing the classification accuracy of the expanded network using a Random Forrest algorithm. This is an interesting approach since the expansion of the network is based on the classification potential of the genes. Instead, Chuang et al. propose the computation of an activity score and then the mutual information score measures the (not necessarily linear) correlation with the class labels. Therefore EgoNet is in some sense more direct and we aim to experiment with it in the future. In general our approach does not necessarily provide the optimum solution. Improvements can be made in the criterion for the augmentation and network expansion so that in addition to the prediction power we consider the biological significance of the genes in a candidate search direction.

Following the network expansion step, we emphasize that the final unified network poses several advantages in detecting new proteins such as protein kinases (AKT1, CDK4, CDK5, CDK6, MAP3K13, PDPK1, PRKACA, PRKCZ, RAF1, STK35, TYK2) and transcription factors (FOSB, FOSL1, HEXIM1, LMO1, LMO3, LMO7, MEOX2, MORF4L2, MYBL2, NR4A1, OPTN, RELA, SETDB1, SMAD3, SMAD4, SMARCC2, SREBF2, TBX15, TCF4, ZFP64) that are implicated in cancer, as well as tumor suppressors (CDKN2A, FANCG, SMAD4, TNFAIP3) and oncogenes (CASC5, DDIT3, EGFR, EWSR1, LMO1, MAFB, RAF1, ZMYM2) by utilizing the Molecular Signatures Database (MSigDB). All these intermediate molecules are essential for the interconnection of higher scoring proteins, a crucial property for the discovery of disease-causing genes [Salhia et al., 2014]. In addition, we confirm that protein-protein interactions, and post-translational modifications (e.g. phosphorylation) can move the activity of a protein from what would have been predicted by its transcription level [Stambuk et al., 2010]. For example, we know that changes in SMAD phosphorylation have been linked to breast cancer metastasis [Chuang et al., 2007].

3.4 Conclusions

We have described a method for exploring the similarities of a set of genes and validating their effectiveness in a two-step process that uses existing biological information. Although more intensive research is needed to investigate the early disease events and to find predictive markers for the course of breast cancer metastatic process, the proposed stepwise refinement approach provides: (i) a meaningful "backbone" Steiner tree and a comprehensive unified network, (ii) putative predictive biomarkers of breast brain metastasis and (iii) a feedback for the validity of the 27 genes of the starting molecular signature as indicative markers for the presence of CTCs.

The assumptions of this work, which were described in the introductory section of this chapter, are that cancer-related genes and their corresponding proteins occupy central locations in biological networks and they are strongly connected with many other genes. We have formulated these assumptions as the solution to the *minimal* Steiner tree problem in graphs, which effectively tries to identify the shortest paths between the genes in our initial list. From the biological point of view there is no clear justification for choosing the shortest paths instead of any other path, except maybe for the "Occam's razor" or "maximum parsimony": the fewest assumptions are often preferred to those positing more. In the next chapter, we are relaxing these assumptions and we consider each gene in our "seed" list as independent module that effectively models its own local biological properties.

Chapter 4

Stacking of network based classifiers

Contents

In the previous chapter we used existing biological knowledge in the form of biological networks in order to identify a single "module" of connected genes around our initial set of 27 genes. The motivation of this strategy is previous research that characterizes cancer biomarkers as "hubs" in the network, that is to say, they tend to encode highly connected proteins [Jonsson and Bates, 2006]. In this chapter instead, we explore another approach: we consider each of the 27 genes as local concentration points of biological functionality and focus on their close "neighbors". Therefore, each biomarker is considered separate from the others rather than part of the same, global functional group. This approach is inline with research that shows disease related genes to occupy peripheral positions in the human interactome [Goh et al., 2007].

The underlying application domain is again the classification task but in this case each of our initial biomarkers is used to construct a "base" network

classifier using its closest neighboring genes. These network classifiers are combined in an "ensemble" to form a two-level classification scheme. At the first level base classifiers are built using the given list of candidate "biomarkers" and the topology of the biological network. In particular, the network structure is taken into account by a search strategy based on random walks for the selection of the genes used in these base-classifiers. At the second level, a meta-classifier is trained to combine in the best possible way the results of the base classifiers. The proposed approach therefore aims to strengthen the classification ability of the initial list of genes and provide more robust generalization guarantees. Our methodology is explained in full detail and promising results in Breast Cancer related scenarios are obtained.

4.1 Introduction

Nowadays the origin and evolution of cancer is conceptually dedicated to an altered network of genetic, epigenetic, metabolic, environmental, and biochemical responses, rather than to a defect of individually molecules. A network-based understanding is essential concerning the biological processes and mechanisms underlying disease progression. Bearing in mind that networks are fundamental in personal and preventive medicine, current research in the field is focused on exploring and elucidating disease networks at the molecular level in order to provide diagnostic, prognostic and predictive biomarkers, and also to intervene to treat disease through the specification of drug targets [Barabási et al., 2011, Vanunu et al., 2010, Vidal et al., 2011].

In the recent years biological networks, either in the form of pathways, gene interaction, or protein interaction networks, have been used to provide insight (explanation) in the research findings or to guide the search strategies for the discovery of "biomarker" genes. In this chapter we propose a novel approach for building predictive models that use the underlying topological information of biological networks in an adaptive way. Starting from the list of candidate biomarkers identified in Chapter 2, we extend the "neighborhood" of each of these genes by taking advantage of the whole graph topology, not only their immediate adjacent genes in the network. This "neighborhood expansion" phase is effectively a "wrapper" based gene subset selection [Kohavi and John, 1997] that uses the classification performance of the neighborhoods as the criterion for further expansion. The selection of subsets of the genes starting from the initial candidate set leads to several gene subsets (one for each initial candidate biomarker) which appear as loosely connected sub-graphs. This observation induces an assumption that each marker gene does not act by itself but rather integrates a number of functionally-neighboring genes towards a specific task or biological action. Thus, these subgraphs or gene subsets are then trained independently to build the same number of classifiers. The objective is then to build a classifier that combines the predictions of these "base" classifiers (i.e. one classifier per biological action initiated by the corresponding biomarker gene) in order to provide the final estimation, in a binary classification problem. The genes selected during the construction of the base classifiers are then considered to form strongly intra-dependent functional groups that can be used as composite features for the unsupervised classification of new samples.

4.2 Methods

The starting point of our work is a list of m genes, $\mathcal{L} = \{g_i, 1 \leqslant i \leqslant m\}$, alongside with a biological network $\mathcal{G} = (V, E)$, where V is the set of vertices corresponding to genes, and E is the list of edges connecting those genes. For the initial set $\mathcal{L}$ of genes we are using the 27 genes derived from our earlier work [Sfakianakis et al., 2014]. These genes were the outcome of the multiple comparisons between normal and cancerous samples in blood and breast tissues. The biological network we use is HINT [Das and Yu, 2012], a protein-protein interaction network that consists, at the time of this writing, of around 11000 nodes (proteins/genes) and 45000 interactions.

The list $\mathcal{L}$ of starting genes represent a candidate "gene signature" for the characterization of a medical condition or a disease such as cancer, or a biological phenotype [Segal et al., 2005, Sotiriou and Piccart, 2007]. We aim at using each of these genes as centers for the construction of independent "classifiers" using the network topology as guidance for the selection of genes that will be included in the classifier built around those centers. The classifiers are then combined into a higher level classifier in order to produce more robust classification performance than the initial list of genes.

Since the final objective is to build a predictive model for the classification of unknown cases (biological samples) into predefined categories (e.g. metastatic versus non-metastatic samples), an additional set of gene expression values $\mathcal{D} = \{(\mathbf{x}_j, y_j), \mathbf{x}_j \in \mathbb{R}^k, y_j = \pm 1, 1 \leqslant j \leqslant N\}$ are used in a supervised learning setting.

Our approach is based on the following two main steps:

- For each gene in the initial list $\mathcal{L}$ we compute their proximities to any other gene in the biological network $\mathcal{G}$.

- The distances between the genes in the initial set $\mathcal{G}$ and any other gene in the network are then used to build a two level hierarchical meta-classifier using the training set $\mathcal{D}$.

4.2.1 Computing gene proximities

The biological network links the different genes and provide a global view of the possible ways that any gene can affect any other gene. In order to quantify such interactions we consider not only the immediate neighbors of a gene in the network, or the shortest path that connects them, but any possible path that contains the genes in question. Using graph theory this problem is formulated as performing random walks in graphs with restart [Chung, 2007, Can et al., 2005].

In more detail the proposed methodology is as follows:

- Taking as input the gene interaction network $\mathcal{G}$, we consider random walks with restart [Lovasz, 1993]: given a parameter β that corresponds to the probability that there's no transition from one node of the graph to any other, the probability distribution for every node in the graph at time $t+1$ is given by:

  $$\mathbf{P}_{t+1} = \beta\mathbf{P}_0 + (1-\beta)\boldsymbol{W}\mathbf{P}_t \tag{4.1}$$

  where $\mathbf{P}_0$ is the initial assignment of weights in the nodes of the graph, and $\boldsymbol{W}$ is the "normalized adjacency matrix" of the graph, i.e.

  $$\boldsymbol{W}_{i,j} = \begin{cases} \frac{1}{\deg(j)} & \text{if node } i \text{ directly interacts with node } j, \\ 0 & \text{otherwise} \end{cases}$$

  where $\deg(j)$ is the degree of a node, that is the number of genes adjacent to it. When the graph is connected then after many transitions the steady state condition $\mathbf{P}_{t+1} = \mathbf{P}_t \equiv \mathbf{P}$ gives:

  $$\mathbf{P} = \beta(\mathbf{I} - (1-\beta)\boldsymbol{W})^{-1}\mathbf{P}_0 \tag{4.2}$$

- In Equation 4.2 the matrix $\beta(\mathbf{I} - (1-\beta)\boldsymbol{W})^{-1} \equiv \mathbf{M}_\beta$ is a stochastic matrix independent of the initial weights $\mathbf{P}_0$ and captures the probability of transitions (in one or more steps) from any node in the graph to any other node (including itself). By taking the negative logarithm of the entries in this matrix we transform it to a matrix of "distances" so that two genes with large transition probability are considered close to each other[1]. This induced gene distance matrix provides the proximity of any gene to any other by taking into account all possible paths in the network as determined by the random walk formulation.

[1] Using the logarithm (instead of only changing the sign, or taking the reciprocal) is not really necessary, but it transforms the values into a more "manageable" range, usually between 1 and 10.

The final result of this step is the determination of the $\mathbf{M}_\beta$ matrix of gene proximities, as explained above. Of course the calculation of this matrix depends on the choice of the parameter β that corresponds to the probability of returning to the initial gene instead of randomly selecting one of the adjacent genes. In the current implementation, the choice of $\beta = 0.4$ has been made.

4.2.2 Building an ensemble of classifiers

The proximity matrix $\mathbf{M}_\beta$ computed in the previous step can be used to select for each $g_i \in \mathcal{L}$ its "nearest" genes, so that we build a different classifier $\mathcal{C}_i$ based on the subset of genes that are most closer to g_i. Subsequently, the different classifiers form a committee or ensemble that if used collectively can increase the classification performance. This is a typical application of ensemble learning [Kittler et al., 1998, Dietterich, 2000] where the output of the different classifiers are combined through voting (majority wins), weighted voting (some classifier has more authority than the others), averaging the results (for regression problems), etc. and there is a large number of approaches taking advantage of Boosting [Schapire, 1990, Freund and Schapire, 1997] or "Bagging" (Boostrap aggregating) [Breiman, 1996] in order to train and build the final classifier.

An alternative way of combining different classifiers was introduced by Wolpert [Wolpert, 1992] under the name of "Stacked Generalization" or "stacking". In this setting the output of many classifiers (or "generalizers" in Wolpert's terminology) form a new training set for another, higher-level, classifier. The basic idea is to train the first-level ("Level 0") classifiers using the original training data set, and then generate a new data set for training the second-level ("Level 1") classifier, where the outputs of the first-level learners are regarded as input features while the original labels are still regarded as labels of the new training data (Figure 4.1).

In our case, we adapt the "stacked generalization" approach as follows:

- Each Level 0 classifier is built from the "neighborhood" of the initial gene list $\mathcal{L}$, using the most proximate genes as determined by the matrix $\mathbf{M}_\beta$. Each base classifier $\mathcal{C}_i$ is trained using Logistic Regression. The rationale for this choice is twofold: First, logistic regression provides a probabilistic output, i.e. a level of confidence that an input sample belongs to the "positive" class (e.g. a "poor prognosis" group) that can be used as a numerical feature for the second level classifier. Secondly, the probabilistic output of a logistic regression classifier is usually well calibrated since it optimizes directly the log loss [Bishop, 2006a]. The choice of how many neighbors to consider for each g_i, which equivalently

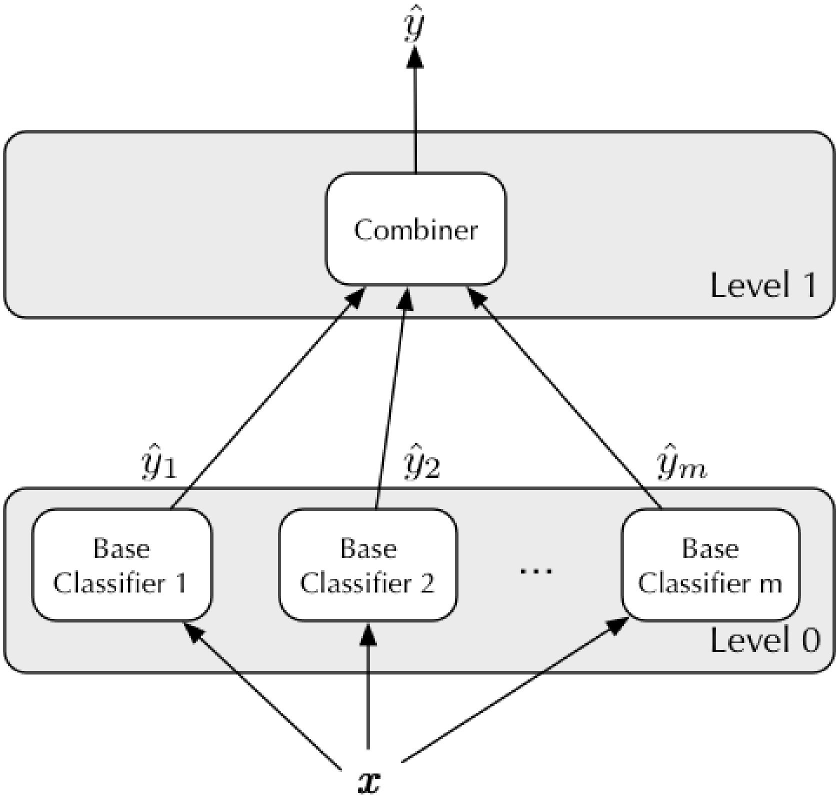


Figure 4.1: Stacked generalization or 2-level "stacking" of classifiers. The classifiers at the first level (Level 0) take as input the input cases and each one of them produces a prediction. The predictions of the first level classifiers are then given as input to the second level (Level 1) classifier ("combiner") that provides the final prediction.

means the radius of the "open ball" [Rosenlicht, 1986] centered at g_i, is determined by an internal cross validation. In this cross validation, increasing values of the radius r, that is the distance from the center gene g_i, are used and the subset of genes contained in the said distance are checked in terms of the impact on classification performance as measured by the Matthews correlation coefficient [Powers, 2011]. The different values of r tried are determined by the desired number of genes to be included in the corresponding ball, for example 3, 5, 10, etc. up to the maximum 30.

- The probabilistic outputs of the base classifiers are combined in a Level 1 Random Forest classifier. As described by Wolpert [Wolpert, 1992], the second level classifier is built using the outputs of the base classifiers in a Leave-One-Out (LOO) scheme. This means that every case in the input training set is kept separate, the base classifiers are trained in the rest of the cases, and their predictions on the test case are then noted.

When every input case has been used as test, a new training set (with the same number of cases as in the initial training set) for the second level classifier has been created based on the predictions of the base classifiers. The random forest classifier is then trained in this new training set. We have chosen the random forest technique as the second level classifier in order to take advantage of its ability to compute "feature importances" as we describe in Section 4.3.3.

The pseudocode of the adaptation of "stacking" and the use of the network information is shown in Algorithm 1.

```
Input: $\mathcal{L}$: list of "seed" genes, $\mathbf{M}_\beta$: gene proximity matrix,
       $\mathcal{D} = \{(\mathbf{x}_j, y_j), \mathbf{x}_j \in \mathbb{R}^k, y_j = \pm 1, 1 \leqslant j \leqslant N\}$: gene expression
       dataset with binary class labels
Output: $\mathcal{C}^0_i$: base classifiers, one for each $g_i$ in $\mathcal{L}$, $\mathcal{C}^1$: second level
        classifier

begin
    foreach $g_i$ in $\mathcal{L}$ do
        find "best" number of neighbors of $g_i$ using $\mathbf{M}_\beta$ and "grid search";
    end
    // List of base classifiers predictions:
    $P \leftarrow \emptyset$;
    for $j = 1$ to $N$ do
        $D_{\text{fold}} = \mathcal{D} - \{\mathbf{x}_j\}$;
        foreach $g_i$ in $\mathcal{L}$ do
            $\mathcal{C}^0_{i,j} \leftarrow$ Fit Logistic Regression with the neighbors of $g_i$ in
            $D_{\text{fold}}$;
            // Predict on the left out sample, and keep
               prediction:
            $P_{i,j} \leftarrow \mathcal{C}^0_{i,j}(\mathbf{x}_j)$;
        end
    end
    // Train 2nd level classifier:
    $\mathcal{C}^1 \leftarrow$ Fit Random Forrest in $P$;
    // Train base level classifiers in the full dataset:
    foreach $g_i$ in $\mathcal{L}$ do
        $\mathcal{C}^0_i \leftarrow$ Fit Logistic Regression with the neighbors of $g_i$ in $\mathcal{D}$;
    end
end
```

Algorithm 1: The adapted "Stacking" algorithm

Table 4.1: Number of sample and characteristics of the public dataset GSE45965

Phenotype	Number of samples
Normal Peripheral Blood	8
Breast Cancer Tumor	50
Circulating Tumor Cells	5
Normal Epithelia	4

4.3 Results and discussion

4.3.1 Data

We have used the recently available data set of [Lang et al., 2015] to test the proposed two-level classification approach. This data set is available as three "sub-series" in the Gene Expression Omnibus (GEO) public database as the "super-series" GSE45965[2]. It consists of 67 gene expression profiles from peripheral blood, circulating cancer, breast cancer tissue, and normal epithelia, as shown in Table 4.1.

For the proper annotation of the gene probes in the data set we have used the UniGene database[3] to perform mappings from the GeneBank identifiers to the Entrez Gene ids and gene symbols. Probes that dont't have a UniGene annotation or that have not been measured in all samples were removed. For the case where multiple probes map to the same Entrez Gene identifier the average (mean) expression value was calculated and used in the downstream analysis. After this annotations and summarization step, we are left with around 15,000 unique genes.

4.3.2 Evaluation

Based on our earlier work [Sfakianakis et al., 2014] presented in Chapter 2, we are interested in the use of computational methods for the profiling of the circulating tumor cells in the blood of breast cancer patients. Therefore we focus on the comparisons between normal peripheral blood (PB) and breast cancer tumors (BC), on one hand, and between CTCs and normal peripheral blood on the other hand. For each binary classification we test Support Vector Machines with Radial Basis Function (Gaussian) kernels and Random Forests type of classifiers alongside with the proposed "stacked" classifiers and a logistic regression using only the candidate biomarkers for the CTC presence

[2] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45965 (accessed February 25, 2016)

[3] http://www.ncbi.nlm.nih.gov/unigene (accessed February 25, 2016)

Table 4.2: Average AUC scores

Classification method	CTC versus PB (with gene selection)	BC versus PB (with gene selection)
Random Forest	0.795 (0.842)	0.917 (0.934)
SVC-RBF	0.980 (0.995)	0.952 (0.995)
Logistic regression with the biomarkers of [Sfakianakis et al., 2014]	0.932	0.979
Subnetwork Stacking	0.960	0.952

identified in [Sfakianakis et al., 2014]. For the evaluation of the performance of the classifiers we are using the area under the curve (AUC) of the Receiver Operating Curve (ROC) [Fawcett, 2006] in order to have a unique metric that takes into account both the true and the false positive rates. In the different comparisons we consider Breast Cancer tumors to form the positive class, unless the comparison contains CTC samples where we consider those cells to be in the positive class. The choice of which is the positive class is not totally irrelevant if combined with the use of AUC as the evaluation criterion since AUC conveys a probabilistic meaning: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [Fawcett, 2006].

For each of the two comparisons we performed a “repeated hold-out” evaluation process where in each iteration a “stratified” split of the randomly shuffled dataset is done. In each iteration the training set consists of the 70% of the samples and the rest 30% is used for testing the trained classifiers. We repeat the same process 100 times and each time we record the AUC achieved by each classifier. The final score of each classification method is the average of the AUC scores achieved in each random split. The results are shown in Table 4.2.

The Table 4.2 contains also the performance results for the Random Forest and the SVM rows when we perform a feature (gene) selection preprocessing step, as follows: From each iteration we keep the most important genes as reported by the Random Forest classifier. The union of all these “important” genes yields 303 genes in the CTC versus peripheral blood comparison and 518 genes in the peripheral blood versus Breast cancer comparison. We then repeat the evaluation in the same “repeated hold-out” process using the same splits but only the most important genes identified previously as features. The results of this second evaluation that is only relevant for the Random Forest and the SVM classifiers are shown in parentheses in Table 4.2.

A first remark on the results shown in Table 4.2 is that all methods exhibit impressive classification performance (above 90%), with the exception of random forests, in the comparison between CTCs and peripheral blood. As expected, the gene selection process improves the performance of the Random Forest and, to a larger extent, the Support Vector Machine classifier (SVC). The candidate biomarkers from our previous work [Sfakianakis et al., 2014] have very good classification power when used as features in logistic regression. This fact emphasizes the importance of the limited number (19) of carefully selected genes from the full set of features in the dataset.

The stacking classifier that we are proposing improves the performance of the set of the 19 genes, but only in the case of the CTCs verus peripheral blood. Of course the base classifiers in the stacking case are not built from the whole set of 19 genes: instead each base classifier is build from the network neighborhood of one of the 19 genes.

4.3.3 Significant neighborhoods

During the training phase each of the base classifiers builds a "neighborhood" around the corresponding seed gene. On the other hand, the second level classifier is a random forest that provides an importance score for its features thus measuring their contribution to the predictive accuracy [Breiman, 2001]. In the proposed stacked generalization methodology the features used are the predictions of the base classifiers and therefore they measure, indirectly, the classification ability of the genes selected in the neighborhood of the corresponding seeding gene.

We take advantage of the iterated hold-out evaluation process so that in each random split of the data into training and testing sets we normalize the importances of the base classifiers based on the performance of the whole 2-level classifier in the test set, as follows:

$$\hat{g}_i^k = \mathrm{AUC}_i * g_i^k$$

where AUC_i is the performance of the stacking classifier in the i-th iteration as measured in terms of the area under the ROC curve, g_i^k is the importance score of the k base classifier as reported by the random forest in the i-th iteration, and $\hat{g}_i^k$ is the corresponding normalized importance of the base classifier. Computing the mean $\hat{g}_i = \sum_{k=1}^{K} \hat{g}_i^k / K$ over all iterations we find the results shown in Fig. 4.2.

In order to see what are the most frequently selected genes for each base classifier we use again the information provided by the multiple splits into training and test sets. We count the number of times each gene is selected in a specific classifier (therefore in the neighborhood of the corresponding

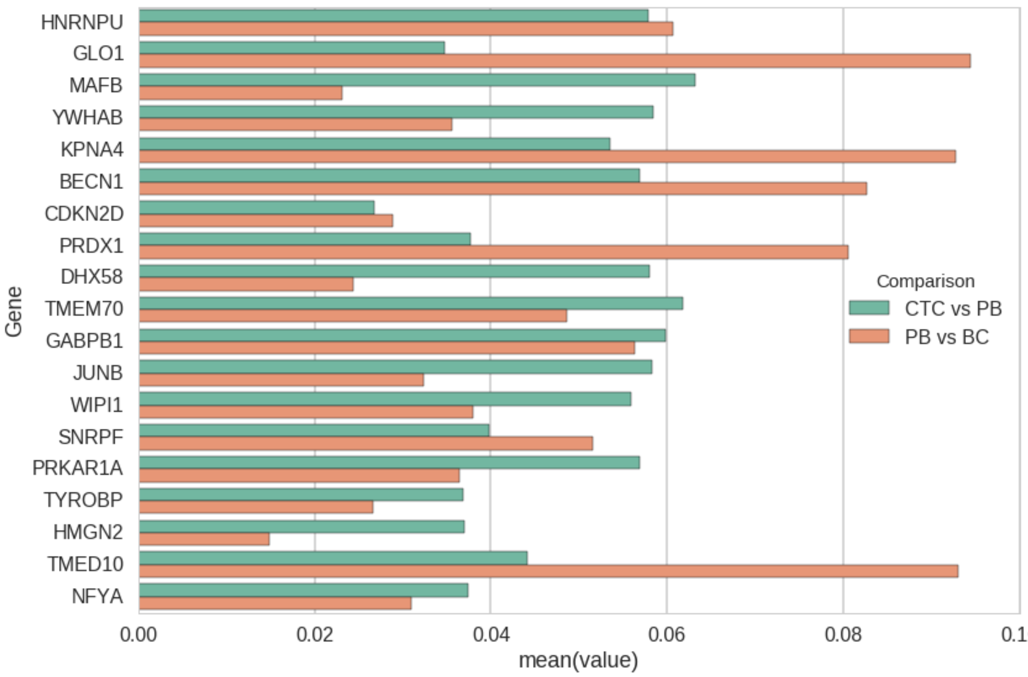


Figure 4.2: The mean importance scores of each base classifier in the different comparisons. Each base classifier is referenced based on the gene that is used as its "epicenter" in the training of the classifier and the selection of the genes most "close" to it.

seed) across all iterations. We then select, for each base classifier, the genes that appear at least in the 60% of the iterations. The results can be seen in Table 4.3 and Table 4.4.

There are some genes that participate in the neighborhoods of more than one seed. This is more evident in the results of the peripheral blood versus breast cancer tumor samples comparison. Also each neighborhood is an "induced" graph of genes where the seeding gene is at the epicenter. In Figures 4.3 and 4.4 we show these induced graphs and how they are connected to each other based on the shared genes. The distance between the seed genes and their neighbors are computed based on the number of the iterations that the neighboring gene was selected during the construction of seed classifier in the repeated hold-out evaluation process. For example, the distance 0.95 means the neighboring gene was selected 95 times out of 100 during this process.

Using the collective set of the genes in the significant of the neighborhoods found in the two comparisons we perform Principal Component Analysis (PCA, see Section 1.4.1.1) and the results can be seen in Figure 4.5. Effectively, we keep only the values for the genes in the two cases separately (156 and 278 for CTC vs PB and PB vs BC, respectively) and then through PCA we keep only

Table 4.3: Genes most commonly selected per seed in the CTC versus PB comparison

Seed/Base classifier	Genes selected in seed's "neighborhood"
BECN1	BCL2, GFI1B, BECN1, ZWINT, PIK3C3
CDKN2D	DAB1, NR4A1, CDKN2D, RXRA, RAB1A, RB1, ASCC2, NEK6, TGFBR1, PSMA1, CDK4, NR4A2, COPS5, INCA1, HSD17B14, CDK6, GRB2, ATXN1, RBPMS, IKZF3
DHX58	ACVR1, POM121, DHX58, KPNA2, SMURF1, SMAD4, TERF2IP, MLH1, ITSN1, APC
GABPB1	CIC, IL16, GABPB1, RSPH14, TRAF2, FANCG, LMO4, GABPA, USO1, SNRPB2
GLO1	SUPT5H, GLO1, BCL10, BIRC2, RIPK1, IKBKB, TNFAIP3, CDCA8, STAT3, NFKBIA, ZDHHC17, TNIP1, CREBBP, GIT2, TAX1BP1, EEF1A1, GTF2E1, TANK, TRIM29, IRAK1
HMGN2	EP300, TERF1, GRB2, TERF2, TINF2, APEX1, TERF2IP, NCK1, XRCC6, HMGN2
HNRNPU	HNRNPU, A1CF, CDKN2A, SYNCRIP, RBM4, HNRNPD, HNRNPH3, HNRNPF, MAPK6, PRMT1
JUNB	JUNB, ATF2, CREB5
KPNA4	MAT2B, HNRNPC, RAC1, KPNA4, KPNA3
MAFB	JUND, CREB5, MAFB, MIS12, ZW10, ZWINT, DDB1, FOS, ATF4, JUN, FOSL1, FOSL2, ZDHHC2, BECN1, ATF2, CEBPG, ATF1, ANAPC5, MAFG, JUNB
NFYA	CSNK2A1, SREBF2, NFYA, ZHX1, NFYC, CDC25A, SP1, APPBP2, NFYB, LUC7L2
PRDX1	TERF1, PRDX1, MYD88, PRDX4, ADH5
PRKAR1A	WNK1, PRKAR1A, UBE2I
SNRPF	SNRPG, LSM6, SNRPE, SNRPF, LSM7
TMED10	RCHY1, JMJD1C, DHODH, AKT1, APOB, TMED2, SNX27, ASS1, TMED10, AKT2
TMEM70	TMEM70, PHC2, SSX2IP
TYROBP	MEOX2, TYROBP, MICA
WIPI1	REN, SETDB1, PPA1, TRIM27, KCTD15, NOTCH2NL, WIPI1, ATXN1, KCTD1, UNC119
YWHAB	YWHAB, YWHAE, RAF1

Each gene neighborhood includes the seed. There are 156 genes in total.

the first 3 principal components with the largest variance as new dimensions. The figures 4.5a and 4.5b show the clear separation of the classes using only the genes in the significant neighborhoods identified in the previous analysis.

4.3.4 Biological Evaluation

We used four databases, the MSigDB[4] (Molecular Signatures Database [Subramanian et al., 2005]), the G2SBC[5] (Genes-to-Systems Breast Cancer [Mosca et al., 2010a]), the DisGeNET[6] (discovery platform on gene-disease associ-

[4] http://software.broadinstitute.org/gsea/msigdb/

[5] http://www.itb.cnr.it/breastcancer/

[6] http://www.disgenet.org/

Table 4.4: Genes most commonly selected per seed in the PB versus Breast Cancer comparison

Seed/Base classifier	Genes selected in seed's "neighborhood"
BECN1	TP53BP2, UVRAG, MAFB, ESR2, MIS12, BAD, NR4A1, MAPK1, BCL2, ZW10, MCL1, ZWINT, CASP3, GFI1B, PIK3R4, MAPRE1, CASP8, BCL2L11, BID, BAK1, BECN1, BNIP3L, ATG14, VDAC1, NSL1, PMAIP1, MAPK8, PPP1CA, PIK3C3, BAX
CDKN2D	DAB1, RB1, CDK6, CDKN2D, RXRA, NEK6, RAB1A, RBPMS, CDK4, TGFBR1, GRB2, NR4A2, COPS5, HSD17B14, IKZF3, PSMA1, ATXN1, ASCC2, INCA1, NR4A1
DHX58	RBPMS, ATP23, ACVR1, AFF4, OBFC1, MTUS2, POMZP3, TRIM23, DHX58, ATP6V1G1, TRAF2, PLSCR1, MDFI, POM121, FAM46C, KPNA2, TANK, SLC25A6, LDLR, RAPGEF3, TRIP6, SMURF1, SMAD4, TERF2IP, NUP54, NMI, MLH1, ITSN1, CCDC85B, APC
GABPB1	CIC, IL16, TRIM27, CDKN2A, CSNK2B, GABPB1, ATXN1, MTUS2, RSPH14, TRAF2, USHBP1, FANCG, LMO4, DAZAP2, QKI, GABPA, PCBP1, ATXN2, USO1, SNRPB2
GLO1	PPP2CA, SUPT5H, GLO1, BCL10, BIRC2, RIPK1, STX11, IKBKB, GCC1, TNFAIP3, CDCA8, STAT3, NFKBIA, ZDHHC17, TNIP1, WWP1, CREBBP, GIT2, TAX1BP1, POLR2E, KRT18, SSX2IP, TARBP2, EEF1A1, SNW1, GTF2E1, TANK, TRIM29, IRAK1, CLIC1
HMGN2	TERF1, XRCC6, TERF2, GRB2, TINF2, APEX1, EP300, NCK1, TERF2IP, HMGN2
HNRNPU	HNRNPU, A1CF, HNRNPF, CDK4, MDM2, PRKCA, CDC45, FN1, CDKN2A, HNRNPA1, SYNCRIP, CDC6, TTR, CRMP1, RBM4, HNRNPD, HNRNPH3, HNRNPUL1, DDX17, CDC5L, MCM5, HNRNPA0, CDK6, MAPK6, DDX5, PRMT1, CHERP, CDC7, HNRNPH1, KAT5
JUNB	FOS, ATF4, CREB5, ATF2, JUNB
KPNA4	RAC1, MAT2B, HNRNPC, KPNA3, KPNA4
MAFB	JUND, CREB5, MAFB, MIS12, ZW10, ZWINT, DDB1, FOS, ATF4, JUN, FOSL1, FOSL2, ZDHHC2, BECN1, ATF2, CEBPG, ATF1, ANAPC5, MAFG, JUNB
NFYA	ESR2, GTF2A2, CSNK2A1, GTF2E2, PCBD1, PWP1, NFYA, SREBF2, CREB1, ZHX1, NFYC, CDC25A, GRB2, SP1, APPBP2, PAPOLG, NFYB, LUC7L2, TBP, SMURF1
PRDX1	TERF1, PRDX1, MYD88, PRDX4, ADH5
PRKAR1A	WNK1, UBE2I, MAPK6, TAB1, PRKAR1A
SNRPF	CLNS1A, SNRPG, LSM6, SNRPE, SMN2, SNRPF, PSMA3, SRSF5, GEMIN2, SRRT, LSM4, LSM7, LSM8, LSM5, SNRPD1, PUF60, SNRPA1, LSM3, SNRPB2, IKZF1
TMED10	STEAP4, RCHY1, APOA1, POFUT1, JMJD1C, DHODH, SH3RF1, TMED10, SORBS2, AKT1, ESR1, PDPK1, APOB, TMED2, NAMPT, SNX27, ASS1, TCL1A, AKT2, AKT1S1, PRKDC
TMEM70	BMI1, SMAD3, KIFC3, SIAH1, L3MBTL3, CSNK2B, TMEM70, GFI1B, TRIM41, FHL3, MFAP1, BYSL, GRB2, KDM1A, FAM161A, NCK1, MAPK6, PHC2, SSX2IP, KAT5
TYROBP	DAB1, RBFOX2, KLRK1, MEOX2, MDFI, ATXN1, DAZAP2, MICA, TYROBP, RBPMS
WIPI1	REN, PPA1, TRIM27, ATXN1, SETDB1, UNC119, KCTD15, NOTCH2NL, WIPI1, KCTD1
YWHAB	BAD, TNFAIP3, PRKCZ, BRAF, RAF1, IGF1R, YAP1, YWHAB, EPB41L3, YWHAE

Each gene neighborhood includes the seed. There are 278 genes in total.

ations [Piñero et al., 2015]), and the FunDO[7] (functional disease ontology

[7] http://fundo.nubic.northwestern.edu/

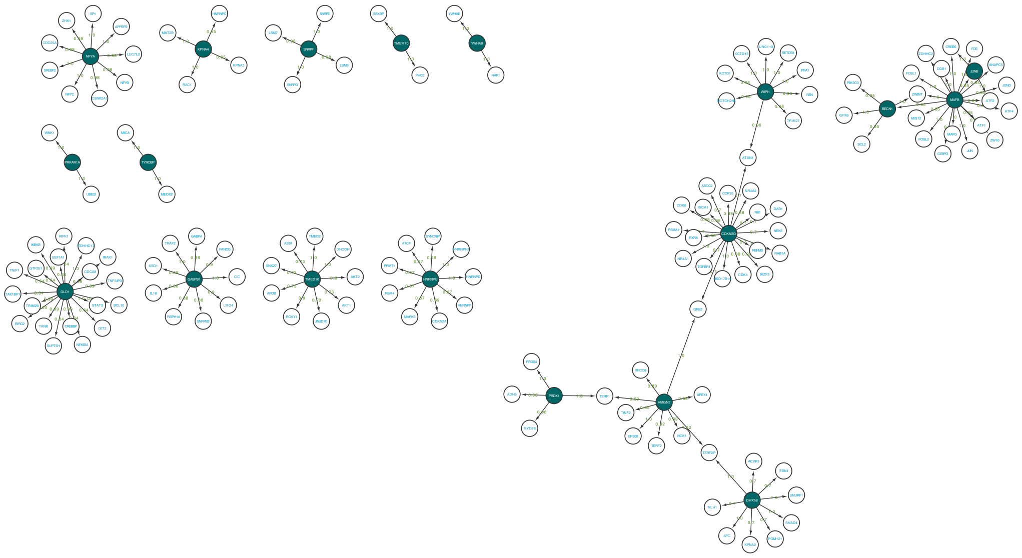


Figure 4.3: The gene neighborhoods that were selected based on the CTC versus Peripheral Blood comparison (Table 4.3). The weights in the edges are calculated based on the number of times the neighboring gene was selected from the base network classifier centered on the corresponding seed gene.

annotations [Osborne et al., 2009]) in order to explore the biological importance of the base-classifiers resulted from the subnetwork stacking classification method. The MSigDB - a collection of annotated gene sets - is used for the membership categorization of base-classifiers by gene families such as oncogenes and tumor suppressors. The G2SBC - a bioinformatics resource for breast cancer study - is used to find the breast cancer-related genes that constitute each base-classifier and also the common molecular alterations (e.g. RNA, DNA, protein) and enriched common pathways. Furthermore, we explored the associations of the 19 seed genes and of the "neighborhood" genes of each base-classifier with cancer, including breast cancer and other cancer types by using one of the most comprehensive resources on gene-disease associations, the DisGeNET, as well as a functional disease ontology (FunDO) annotations database. PANTHER database[8] [Mi and Thomas, 2009] is used for post-assessment of the statistical over-representation of our large gene lists. In addition, we take advantage of the recent study of [Lang et al., 2015] regarding the expression profiling of CTCs in metastatic breast cancer in order to evaluate the "neighborhood" genes of the base-classifiers.

The biological results are as follows:

1. Oncogenes and tumor suppressors: A significant number of oncogenes

[8] http://pantherdb.org/

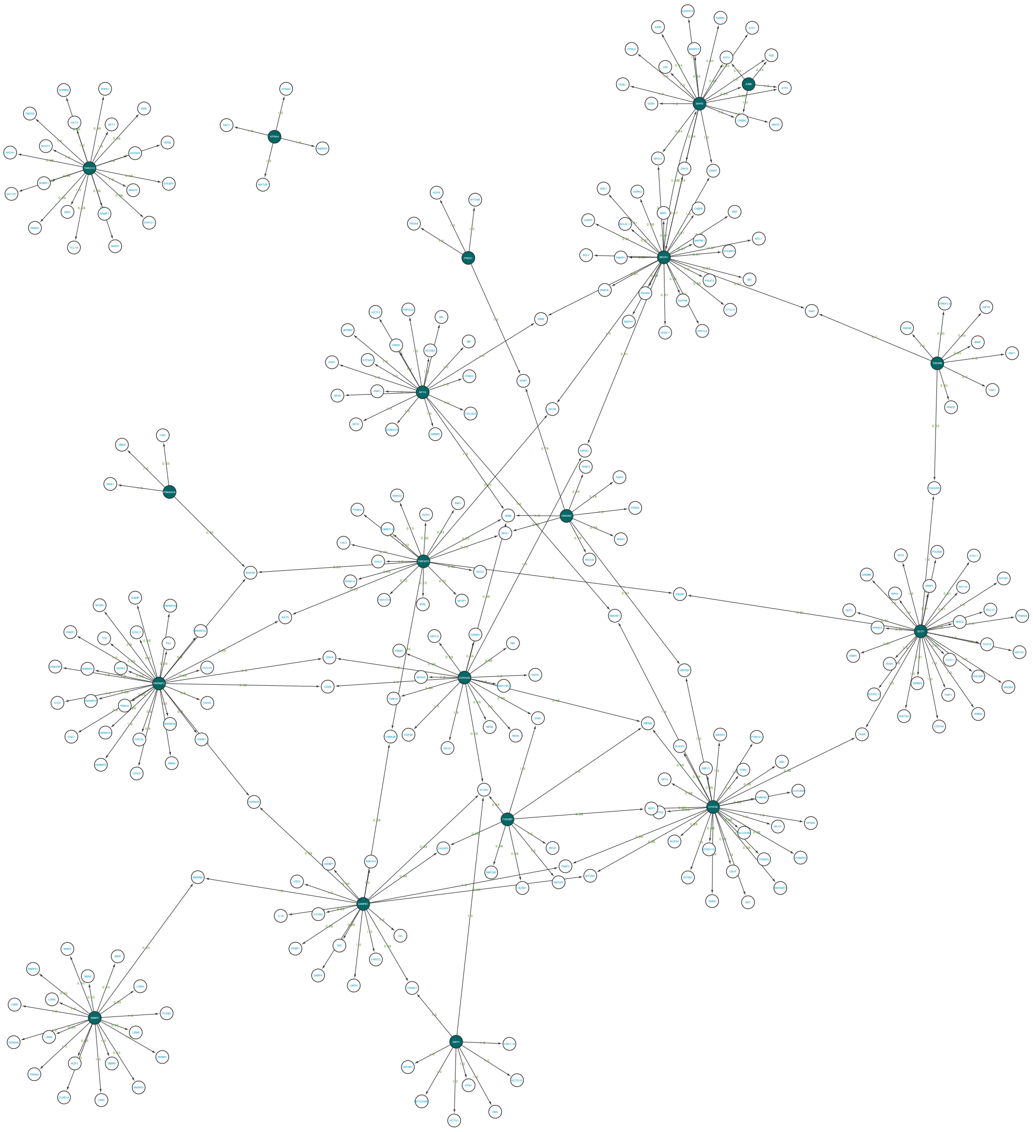
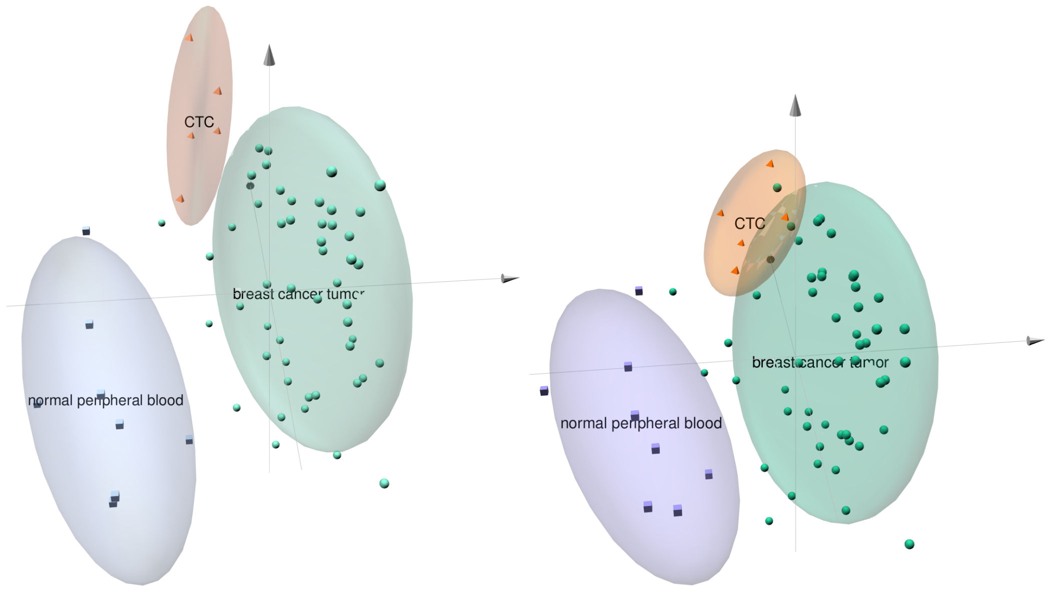


Figure 4.4: The gene neighborhoods that were selected based on the Peripheral Blood versus Breast Cancer tumor comparison (Table 4.4). The weights in the edges are calculated based on the number of times the neighboring gene was selected from the base network classifier centered on the corresponding seed gene.

(e.g. AKT1, JUN, CDK4, CREB1), and tumor suppressor genes (e.g. TNFAIP3, SMAD4, CDKN2A) with known breast cancer alterations, as well as translocated cancer genes (e.g. CIC, MAFB) and transcription factors (e.g. CREB5, FOS, ESR1) are present in base-classifiers and play a key role in the interconnection of expression-responsive genes (e.g. JMJD1C, IL16, FOSL2, JUNB, JUND, TYROBP [Chuang et al., 2007, Lang et al., 2015] (Supplementary Table S1).



(a) CTC vs PB comparison

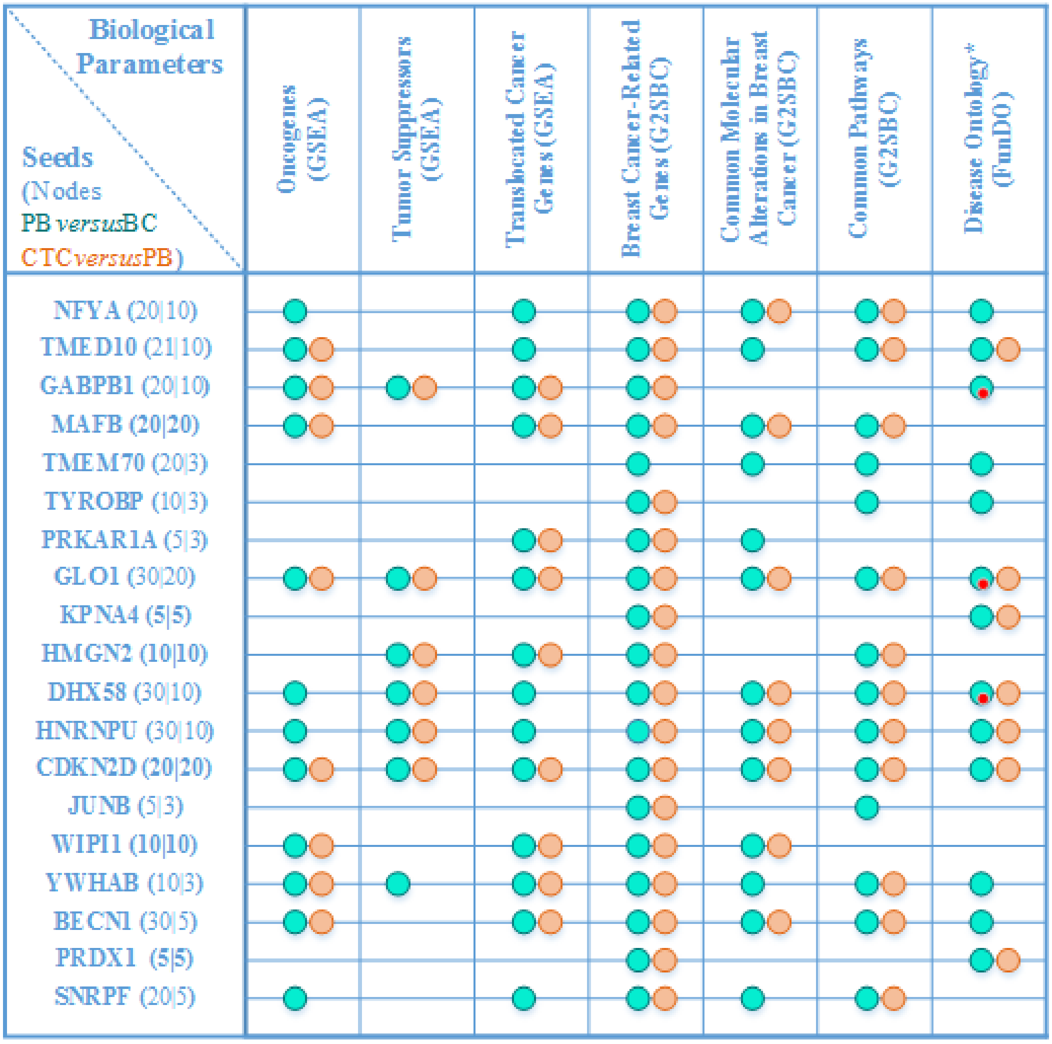
(b) PB vs BC comparison

Figure 4.5: A three dimensional "PCA plot" using the genes in the neighborhoods of the seeds that were selected based on the CTC versus Peripheral Blood comparison (a), and in the Peripheral Blood versus Breast Cancer comparison (b).

2. Breast Cancer-related genes: A number of genes that are associated with breast cancer (e.g. SREBF2, BECN1, GRB2, MAPK6, CASP3) participate in the "neighborhood" of each base-classifier. Some of these genes provide common molecular alterations (e.g. RNA, DNA, protein) and enriched common pathways.

3. Expression-responsive genes: As shown in Table 4.5, a number of genes that constitute some of the base-classifiers are detected as differentially expressed in CTC *versus* PB and/or CTC *versus* BC, according to the work of [Lang et al., 2015]. Seven seed genes (TMED10, TYROBP, JUNB, GLO1, PRKAR1A, HNRNPU, PRDX) that function as "epicenter" of their corresponding base-classifiers are also detected as differentially expressed in CTC *versus* PB and/ or CTC *versus* BC [Lang et al., 2015] validating our initial results [Sfakianakis et al., 2014].

In Figure 4.6 gene family characteristics and breast cancer-related properties are summarized, as obtained through the MSigDB, G2SBC and FunDO databases. As shown, the GABPB1, GLO1 and DHX58 modules (seed subnetworks) in PB *versus* BC comparison provide the most compact biological information and might be used more effectively as a non-invasive way to monitor and prevent metastasis.



Biological Parameters / Seeds (Nodes PB *versus* BC CTC *versus* PB)	Oncogenes (GSEA)	Tumor Suppressors (GSEA)	Translocated Cancer Genes (GSEA)	Breast Cancer-Related Genes (G2SBC)	Common Molecular Alterations in Breast Cancer (G2SBC)	Common Pathways (G2SBC)	Disease Ontology* (FunDO)
NFYA (20|10)	🔵		🔵	🔵🟠	🔵🟠	🔵🟠	🔵
TMED10 (21|10)	🔵🟠		🔵	🔵🟠	🔵	🔵🟠	🔵🟠
GABPB1 (20|10)	🔵🟠	🔵🟠	🔵🟠	🔵🟠			◉
MAFB (**20|20**)	🔵🟠		🔵🟠	🔵🟠	🔵🟠	🔵🟠	
TMEM70 (20|3)				🔵	🔵	🔵	🔵
TYROBP (10|3)				🔵🟠		🔵	🔵
PRKAR1A (5|3)			🔵🟠	🔵🟠	🔵		
GLO1 (30|20)	🔵🟠	🔵🟠	🔵🟠	🔵🟠	🔵🟠	🔵🟠	◉🟠
KPNA4 (**5|5**)				🔵🟠			🔵🟠
HMGN2 (**10|10**)		🔵🟠	🔵🟠	🔵🟠		🔵🟠	
DHX58 (30|10)	🔵	🔵🟠	🔵	🔵🟠	🔵🟠	🔵🟠	◉🟠
HNRNPU (30|10)	🔵	🔵🟠	🔵	🔵🟠	🔵🟠	🔵🟠	🔵🟠
CDKN2D (**20|20**)	🔵🟠	🔵🟠	🔵🟠	🔵🟠	🔵🟠	🔵🟠	🔵🟠
JUNB (5|3)				🔵🟠		🔵	
WIPI1 (**10|10**)	🔵🟠		🔵🟠	🔵🟠	🔵🟠		
YWHAB (10|3)	🔵🟠	🔵	🔵🟠	🔵🟠	🔵	🔵🟠	🔵
BECN1 (30|5)	🔵🟠		🔵🟠	🔵🟠	🔵🟠	🔵🟠	🔵
PRDX1 (**5|5**)				🔵🟠			🔵🟠
SNRPF (20|5)	🔵		🔵	🔵🟠	🔵	🔵🟠	

Figure 4.6: Simplified representation of the biological information included within each seed-subnetwork (module) disclosed in the "PB versus BC" and "CTC versus PB" comparisons. The presence of biological information in each module of both comparisons is highlighted with turquoise (PB versus BC) and brown (CTC versus PB) comparisons regarding seven selected parameters. The numbers in parentheses are the number of nodes in "PB versus BC" and "CTC versus PB" comparisons. Identical nodes in both comparisons are in bold. In the last column the cancer association found only by FunDO is shown. The red dots inside the circles (GABPB1, GLO1, DHX58) highlight the functional association of the module-genes with breast cancer.

By searching the DisGeNET database we discovered cancer and/or breast cancers associations (mammary neoplasms, ductal carcinoma, inflammatory breast carcinoma, secondary malignant neoplasm of bone etc.) or phenotype associations (solid tumor, carcinogenesis, inflammation, tumor progression, tumor angiogenesis, tumor expansion, neoplasm metastasis, alkaline phosphatase adverse event etc) with most of the 19 seed genes and a Disease Specificity Index ranging from 0.39 to 0.88. In addition, by using FunDO we revealed the functional associations of cancer, and/or breast cancer and/or other cancer types with the "neighborhood" genes of the base-classifiers.

Overall, the above biological findings provide insight into the nature of the "neighborhood" genes of the base-classifiers and their interactions. A number of breast cancer-related genes are to a greater or lesser degree interconnected with a) family members of oncogenes, tumor suppressors, transcription factors, protein kinases etc, and b) expression-responsive genes that can be involved in disease in many different ways (Table 4.5). Furthermore, DisGeNET and FunDO provide additional information about the cancer relations and cancer-related phenotype linkage with the seed genes, as well as the functional associations of the interconnected genes/proteins (modules) with cancer. Since each module entails distinct biological properties, several module-specific assumptions can be tested on groups of genes regarding their significance in a CTC subpopulation and their metastatic potential in a distinct homogeneous population of patients (same tumor type and disease stage), providing reliable markers for disease prognosis, treatment response and the overall clinical outcome of patients. Such hypothesis could be validated either computationally in appropriate publicly available datasets or experimentally in targeted clinical studies.

As questions to the meaning of the disease associations and the molecular events driving breast cancer pathology remain, we support the notion that the proposed stacking classifier generates highly informative base-classifiers and achieves higher discrimination and prediction potential than the set of the 19 genes.

4.4 Conclusions

Here, we present abstract elements of a wealth of biological information that has enabled us to confirm our first assumptions concerning biomarkers indicative of the CTC profiles, but also our multilevel methodological approaches.

In the present chapter we have described a method for combining biological information in the form of networks in a "ensemble"-based classification scheme. Biological networks are used in order to reveal gene interactions based on the whole set of possible ways (paths) that two genes "communicate" instead of just considering their immediate neighbors. We then consider a set of input genes that are used as seeds for building a corresponding set of "base classifiers" based on the direct or indirect interactions the seeds have with other genes. As input genes we have used the list of candidate biomarkers that were the

Table 4.5: Expression-Responsive Genes – Results of the comparisons

	PB *versus* BC			CTC *versus* PB	
Seeds	Nodes	Differential expression in [Lang et al., 2015]*	Seeds	Nodes	Differential expression in [Lang et al., 2015]†
NFYA	20	GTF2E2, PCBD1	NFYA	10	–
TMED10	21	ASS1, TMED10, TMED2	TMED10	10	JMJD1C, TMED10
GABPB1	20	DAZAP2, IL16	GABPB1	10	FANCG, IL16
MAFB	20	FOSL2	MAFB	20	FOSL2, JUNB, JUND
TMEM70	20	–	TMEM70	3	PHC2
TYROBP	10	DAZAP2, TYROBP	TYROBP	3	TYROBP
PRKAR1A	5	PRKAR1A	PRKAR1A	3	PRKAR1A
GLO1	30	IRAK1, GLO1	GLO1	20	CREBBP
KPNA4	5	–	KPNA4	5	MAT2B
HMGN2	10	–	HMGN2	10	–
DHX58	30	KPNA2, TRIP6, SLC25A6, MLH1, ATP6V1G1	DHX58	10	–
HNRNPU	30	CDK6, HNRNPU, RBM4, HNRNPA1, FN1	HNRNPU	10	RBM4
CDKN2D	20	CDK6	CDKN2D	20	ATXN1
JUNB	5	–	JUNB	3	JUNB
WIPI1	10	–	WIPI1	10	ATXN1, NOTCH2NL
YWHAB	10	RAF1, YWHAE	YWHAB	3	RAF1
BECN1	30	VDAC1, ATG14, MAPRE1, PPP1CA, MAPK1	BECN1	5	BCL2
PRDX1	5	PRDX1	PRDX1	5	MYD88
SNRPF	20	SNRPE, LSM3, SRSF5, LSM5	SNRPF	5	–

Each seed-subnetwork (module) includes a number of genes that are detected as differentially expressed in "CTC *versus* PB" and/or "CTC *versus* BC" according to the work of [Lang et al., 2015]. Seven seed genes are also identified as expression-responsive genes. Abbreviations: CTC, circulating tumor cells; PB, peripheral blood; BC, breast cancer.

* Differential expressed genes in the comparison "CTC *versus* Tumor" of [Lang et al., 2015].

† Differential expressed genes in the comparison "CTC *versus* Peripheral Blood" of [Lang et al., 2015].

output of the statistical analysis presented in Chapter 2 and each of these seeds expands to a "neighborhood" of strongly communicating, direct or indirect, neighbors based on the random walks in the underlying biological network. The induced neighborhoods are built independent from each other, contrary to the methodology we followed in the previous chapter. Each neighborhood built around the corresponding seed/input gene is then used to build a classifier that takes advantage of the classification power of only the genes contained in the neighborhood, effectively making local decisions based on the biological functions of the neighbors involved. The base classifiers are subsequently used to train a second level classifier that can potentially combine intelligently the successes and failures of them.

The use of biological networks and random walks in graphs have been studied in multiple publications [Shi et al., 2012, Leiserson et al., 2014, Wang et al., 2014, Hofree et al., 2013]. Especially HotNet2 [Leiserson et al., 2014] uses similar random walk-based schemes, but our aim is to construct an adaptive, gene signature initialized, and biologically driven classifier. The results are indeed promising and make stronger the selection of our initial gene list used for the initialization, but further evaluation, tuning, and validation are needed. The implementation is based on the “scikit-learn” machine learning framework for Python [Pedregosa et al., 2011] and all the code and the data can be found at `https://github.com/sgsfak/subnet_stacking`.

In the next chapter, we build upon the results of this chapter by exploring the discriminating abilities of the “neighborhoods” built around our initial set of genes in an “unsupervised learning” setting.

Chapter 5

A biology-adapted Gaussians Mixture Model

Contents

In the previous chapter we have used the 27-genes signature and prior biological information in order to build computational tools for the classification task. In this chapter, we are focusing on the clustering task, that is the categorization of samples into a number of groups so that samples grouped together are "similar" to each other. The approach we use is largely based on the Gaussians (Normal) Mixture Model (GMM) that provides a mathematically appealing and extremely flexible model for this task. On the other hand, the intricacies of the domain, such as the high dimensionality and noise, present real challenges for the employment of finite mixture models in bioinformatics. Here, we introduce a biology-driven adaptation of GMM where information from biological networks and similar metadata sources, such as Gene Ontology, *can be used to constrain the model. The adapted methodology is generic and we present two test cases. First, we evaluate it using information extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. Additionally, we use the neighborhoods of the 27-genes extracted as described in Chapter 4 to parameterize the adapted GMM and report its results.*

5.1 Introduction

In Chapter 1 we introduced the problems of high dimensionality and the inherent noise in the analysis of high-throughput data produced by technologies such as gene expression microarrays. Indeed, in the case where the number of features is much larger than the number of observations standard statistical methods are either completely inappropriate or induce a high variance and overfitting. As mentioned in Section 1.4.1.1, possible ways to deal with the "few samples, many features" situation include techniques like feature selection [Saeys et al., 2007] and regularization or shrinkage methods (e.g. [Tibshirani, 1996, Tibshirani et al., 2002, Zou and Hastie, 2005]).

In this chapter we instead focus on the integration of domain specific knowledge with the statistical learning methods to address some of challenges mentioned above. In particular, we aim at taking advantage of the known functional relationships of certain genes, such as their annotations in the Gene Ontology [Ashburner et al., 2000] or the KEGG [Kanehisa and Goto, 2000] pathways they participate in, to infer better probabilistic models for their expression. The problem we aim to attack is the *clustering* (or "unsupervised classification") of patients samples i.e. to group them in unknown target categories, instead of the classification task we pursued in the previous chapter. The underlying framework is based on finite mixtures of Gaussian distributions modified to account for the information originating from the molecular biology.

The rest of the chapter is structured as follows. First, we introduce the methodology as an adaptation of the *model based clustering* and we describe a modified "Expectation Maximization" algorithm to search for possible solutions. Next, we perform a number of experiments using public datasets and information from some well known gene networks and compare the results with other clustering algorithms. Finally in Section 5.3.3 we concentrate in the neighborhoods of the 27 genes that we have singled out as described in Chapters 2 and 4. We are using these neighborhoods in the modified model based framework and present their performance in the unsupervised classification tasks.

5.2 Methods

5.2.1 Finite Mixture Models

Mixture models [McLachlan and Peel, 2000] present a probabilistic framework both for building complex probability distributions (e.g. *density estimation*) as linear combinations of simpler ones but also for clustering data, a task also known as *unsupervised learning*. Assuming that our data consists of N observations $\{\mathbf{x}_j\}$, the probability density function of a random sample under a g-component mixture model is given as

$$f(\mathbf{x}_j;\boldsymbol{\Theta}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{x}_j;\boldsymbol{\theta_i}) \tag{5.1}$$

where $\boldsymbol{\Theta}$ is the collection of the unknown parameters π_i that are usually referred as "mixing coefficients", and $\boldsymbol{\theta_i}$, which are the parameters of the component densities f_i. In order to make (5.1) a proper density function the following constraints are also imposed: $0 \leqslant \pi_i \leqslant 1$ and $\sum_{i=1}^{g} \pi_i = 1$.

An important specialization of (5.1) is the Gaussian Mixture Model (GMM) where the parametric family of the component density is assumed to be the Gaussian distribution but with different means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$:

$$\begin{aligned} f_i(\mathbf{x}_j;\boldsymbol{\theta}_i) &= \mathcal{N}(\mathbf{x}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) \\ &\equiv \frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma}_i|}} e^{-\frac{1}{2}(\mathbf{x}_j-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_j-\boldsymbol{\mu}_i)} \end{aligned} \tag{5.2}$$

where p is the dimensionality of the sample vectors (i.e. the number of genes).

5.2.2 Integrating biological knowledge

In the proposed model we assume that genes can be classified in K functional groups, using GO Biological Processes for example. The fundamental assumption of this model is that genes that belong to different groups are independent whereas in the same group the gene relationships are unconstrained. Additionally genes that do not belong to any functional group are modeled as totally independent random variables. Since for the Gaussian distribution independence is equivalent to uncorrelatedness the proposed model implies the following structure for the covariance matrix

$$\widetilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Sigma}^{(1)} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}^{(2)} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}^{(K)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{D}^{(r)} \end{bmatrix} \tag{5.3}$$

where each of $\mathbf{\Sigma}^{(k)}$ is the (unconstrained) covariance (sub)matrix for the genes belonging to the k group, and $\mathbf{D}$ is the diagonal covariance matrix of the r genes that do not belong to any group. The covariance model shown above is related to the covariance matrix estimation, and in particular estimating **sparse** covariance matrices, e.g. as in [Banerjee and El Ghaoui, 2008].

The structure of (5.3) is imposed on every component of the mixture model so that (5.2) is rewritten as

$$f_i(\mathbf{x}_j;\boldsymbol{\theta}_{\mathbf{i}}) = \mathcal{N}(\mathbf{x}_j;\boldsymbol{\mu}_i,\widetilde{\mathbf{\Sigma}}_i) \tag{5.4}$$

and then taking into account the block diagonal structure of (5.3) the normal density of (5.4) factorizes into

$$f_i(\mathbf{x}_j;\boldsymbol{\mu}_i,\widetilde{\mathbf{\Sigma}}_i) = \mathcal{N}(\mathbf{x}_j^{(r)};\boldsymbol{\mu}_i^{(r)},\mathbf{D}_i^{(r)})\prod_{k=1}^{K}\mathcal{N}(\mathbf{x}_j^{(k)};\boldsymbol{\mu}_i^{(k)},\mathbf{\Sigma}_i^{(k)}) \tag{5.5}$$

where we have used the "exponent" $^{(k)}$ to refer to the selection of the genes (and means and covariance sub-matrices) belonging to the k category or the rest genes.

Now if we take a mixture of Gaussians of the form (5.5) the equation (5.1) becomes

$$\begin{aligned} f(\mathbf{x}_j;\mathbf{\Theta}) &= \sum_{i=1}^{g}\pi_i\mathcal{N}(\mathbf{x}_j^{(r)};\boldsymbol{\mu}_i^{(r)},\mathbf{D}_i^{(r)}) \\ &\cdot\prod_{k=1}^{K}\mathcal{N}(\mathbf{x}_j^{(k)};\boldsymbol{\mu}_i^{(k)},\mathbf{\Sigma}_i^{(k)}) \qquad (5.6)\\ &= \sum_{i=1}^{g}\pi_i\prod_{k=1}^{K+1}\mathcal{N}(\mathbf{x}_j^{(k)};\boldsymbol{\mu}_i^{(k)},\mathbf{\Sigma}_i^{(k)}) \end{aligned}$$

with $\mathbf{\Sigma}_i^{(K+1)} \equiv \mathbf{D}_i^{(r)}$.

5.2.3 The Expectation-Maximization algorithm

The Expectation-Maximization (EM) algorithm [McLachlan and Krishnan, 1997] is useful in cases where we want to find maximum likelihood solutions for models that have hidden variables. In some cases these hidden variables are introduced on purpose in order to simplify the maximum likelihood estimations of the model's parameters [Bishop, 2006b]. For the problem at hand a maximum likelihood estimation of the parameters of the mixture model can be performed by the introduction of the "missing" (or unobserved) data $\mathbf{z}_j, 1 \leqslant j \leqslant N$, where

$\mathbf{z}_j = (z_{j1} \ldots z_{jg})$ is defined as

$$z_{jk} = \begin{cases} 1 & \text{if } \mathbf{x}_j \text{ was generated by component } k \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

Using the Bayes rule we can compute the support (or "responsibility") each sample provides to a given component density as the conditional probability

$$\begin{aligned} \tau_{ji} \equiv \Pr(z_{ji} = 1 | \mathbf{x}_j; \boldsymbol{\Theta}) &= \frac{\Pr(z_{ji} = 1) f_i(\mathbf{x}_j; \boldsymbol{\theta}_i)}{\sum_{c=1}^{g} \Pr(z_{jc} = 1) f_c(\mathbf{x}_j; \boldsymbol{\theta}_c))} \\ &= \frac{\pi_j f_i(\mathbf{x}_j; \boldsymbol{\theta}_i))}{\sum_{c=1}^{g} \pi_c f_c(\mathbf{x}_j; \boldsymbol{\theta}_c))} \end{aligned} \tag{5.8}$$

The EM algorithm operates iteratively in two stages. In the E-step the estimations of the "missing" data $\mathbf{z_j}$ using τ_{ji} are computed based on the current estimation of the parameter values and the observed data. In the M-step the estimations of the τ_{ji} in the E-step are used in order to update the estimations of the model parameters $\boldsymbol{\Theta} = \{\pi_1, \ldots, \pi_g, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g\}$.

Using the biological knowledge and the sparse structure of the components' covariance matrix shown in (5.3) in the Gaussians mixture model of (5.6), it is relatively easy to show that

- In the E-step the "responsibilities" are updated based on the current model parameters as

$$\tau_{ji} = \frac{\pi_i^{\text{cur}} \prod_{k=1}^{K+1} \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_i^{(k),\text{cur}}, \boldsymbol{\Sigma}_i^{(k),\text{cur}})}{\sum_{s=1}^{g} \pi_s^{\text{cur}} \prod_{k=1}^{K+1} \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_s^{(k),\text{cur}}, \boldsymbol{\Sigma}_s^{(k),\text{cur}})} \tag{5.9}$$

- In the M-step the new model parameters can be separately computed per functional group as

$$\boldsymbol{\mu}_i^{(k)} = \frac{\sum_{j=1}^{N} \tau_{ji} \mathbf{x}_j^{(k)}}{\sum_{j=1}^{N} \tau_{ji}} \tag{5.10}$$

$$\boldsymbol{\Sigma}_i^{(k)} = \frac{\sum_{j=1}^{N} \tau_{ji} (\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)})(\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)})^T}{\sum_{j=1}^{N} \tau_{ji}} \tag{5.11}$$

$$\pi_i = \frac{\sum_{j=1}^{N} \tau_{ji}}{N} \tag{5.12}$$

Appendix B contains the full details for this resolution.

5.3 Results

5.3.1 Implementation

We have implemented the proposed model and in this section we present some preliminary results. The implementation of the algorithm follows the standard EM for Gaussian mixtures (e.g. see [Bishop, 2006b], Section 9.2.2) but uses the (5.9) and (5.10, 5.11, 5.12) in the E and M-steps respectively.

The sparse structure of the covariance matrix (5.3) allows a dimensionality reduction since the functional groups, and subsequently the sub-matrices $\mathbf{\Sigma}^{(k)}$ along the diagonal, include a significantly less number of genes than the original data set. Nevertheless, it can still be the case that these numbers of genes are still bigger than the number of samples, resulting in rank deficiencies in the estimation of $\mathbf{\Sigma}^{(k)}$ according to (5.11). For this reason, in these cases, as a "hard" imposed dimensionality reduction, we compute a rank truncated estimation of $\mathbf{\Sigma}^{(k)}$ using its Singular Value Decomposition (SVD)[1].

As described in [McLachlan and Peel, 2000] and elsewhere, the EM can have a slow convergence and even more it can be "trapped" in a local maximum of the likelihood function. Therefore multiple executions of the algorithm beginning from randomly selected initial values is usually recommended.

5.3.2 Evaluation using common biological networks

In order to perform some evaluation of our method two data sets are used:

- A Breast Cancer data set [Huang et al., 2003] where there exist 52 samples with 18 samples exhibit recurrence of tumor and 34 do not.
- A Prostate Cancer data set [Singh et al., 2002] where there exist 52 tumor samples and 50 normal samples.

Both are based on the Affymetrix HG-U95Av2 chip, containing 12625 probesets that we have preprocessed using the GC-RMA probe normalization and summarization method. Furthermore, genes that exhibit low variation or do not have an Entrez Gene identifier were filtered.

We have selected a number of KEGG "pathways" to be used as functional groups. These pathways are shown in Table 5.1. More or less all of these pathways are examples of Gene Regulatory Networks, which are essentially graphs with genes as vertices and edges that represent regulation (e.g. activation, inhibition) of gene expression. Here we ignore the internal structure of

[1] This is of course equivalent to the Principal Component Analysis (PCA) technique, see Appendix A.

Table 5.1: The KEGG pathways used in the tests

	Pathway id	Pathway name
1	04115	p53 signaling pathway
2	04210	Apoptosis
3	04370	VEGF signaling pathway
4	05010	Alzheimer's disease
5	05012	Parkinson's disease
6	05014	Amyotrophic lateral sclerosis (ALS)
7	05016	Huntington's disease
8	05200	Pathways in cancer
9	05210	Colorectal cancer
10	05212	Pancreatic cancer
11	05213	Endometrial cancer
12	05215	Prostate cancer
13	05222	Small cell lung cancer
14	05223	Non-small cell lung cancer
15	05416	Viral myocarditis

these graphs and consider them as just “modules of biological functionality” so that genes participating in the same pathway are considered probabilistically dependent.

The EM algorithm estimates the parameters of the mixture model (5.1) but in the process it computes the “support” (5.8) each sample provides for each cluster and based on these values a “hard” clustering can be done by assigning each sample to the cluster it mostly supports. But in order to really evaluate the performance of our approach, the information about the true underlying clusters is needed. Unfortunately, this is not possible for the test data sets described above, since they are real and their biological underpinnings are not fully known.

A possible approach is to use the samples' phenotypes as yardsticks in this evaluation. So, since in both test data sets a binary classification is possible we request the identification of 2 clusters, i.e. our modified EM is run with $g = 2$ component mixture model. In comparison to our method we include the clusterings performed by two well known algorithms: K-means and PAM, which is a robust version of k-means based on “medoids” (see Section 1.3.2. Due to the high dimensionality we weren't able to get results for other EM and model based clustering approaches as implemented in the MCLUST software [Fraley and Raftery, 1999].

Table 5.2: BHI Results

Algorithm	BHI Breast Cancer	BHI Prostate Cancer
kmeans	0.55	0.52
pam	0.56	0.51
our EM	0.56	0.49

Based on class labels of the samples (e.g. "relapse" vs. "non-relapse") we can use the Biological Homogeneity Index (BHI [Datta and Datta, 2006]) to check whether the clusters produced by the different clustering algorithms are indeed homogeneous. Ideally, if, for example, all the tumor samples are in the same cluster and all the normal samples are in the other one, BHI will be 1, which is its maximum value. The formula for BHI is given in (5.13).

$$\mathrm{BHI} = \frac{1}{g}\sum_{i=1}^{g}\frac{1}{N_i(N_i-1)}\sum_{\substack{x\neq y\\ x,y\in\mathcal{D}_i}} \mathbb{I}(C(x)=C(y)) \tag{5.13}$$

So basically for any pair x, y of different samples clustered together in a cluster $\mathcal{D}_i$ that "contains" N_i samples, we check whether they have the same class label by the indicator function $\mathbb{I}(C(x) = C(y))$. This is done for every distinct pair of samples in every cluster and the number of matching class labels is properly normalized.

The results of the BHI measure for the two data sets and the three clustering algorithms are shown in Table 5.2. We see that more or less all the algorithms exhibit the same performance in terms of phenotype homogeneity. In these tests our EM was initialized based on the results of the K-means but in the one data set homogeneity was improved whereas in the other was worsened.

In order to get a better understanding on the clustering results, in Table 5.3 and Table 5.4 we show the proportion of each class (i.e. Relapse/Non-Relapse, Normal/Tumor) in each of the clusters identified by the algorithms. The identification of the clusters along the different algorithms has been done based on the Euclidean distances of the "prototypes" of the algorithms, i.e. the centers of the K-means, the medoids of PAM, and the mean vectors of EM. We then use "majority vote" in each cluster to classify the members of the cluster to the most frequent class label. The "winning" class is shown with a boldface in the tables. The results of this "classification" task, for each algorithm and across all clusters, in terms of misclassification rate, sensitivity, and specificity are also

Table 5.3: Classification results (Breast Cancer)

Algorithm	Clusters # 1	 # 2	Miscl. rate	Sensitivity	Specificity
kmeans	**12**/11	**22**/7	0.346	0	1
PAM	12/**12**	**22**/6	0.346	0.667	0.647
our EM	**14**/12	**20**/6	0.346	0	1

Table 5.4: Classification results (Prostate Cancer)

Algorithm	Clusters # 1	 # 2	Miscl. rate	Sensitivity	Specificity
kmeans	**19**/10	31/**42**	0.402	0.808	0.380
PAM	**21**/12	29/**40**	0.402	0.769	0.420
our EM	**22**/18	28/**34**	0.451	0.654	0.440

shown. Of course the results are poor in terms of classification performance but they nevertheless show how the different algorithms separate the data.

5.3.3 Evaluation using the 27-genes and their "neighbors"

As demonstrated in the previous paragraph the proposed EM algorithm with block-diagonal covariance matrices in the Gaussian components does not appear to yield biologically relevant clusters when generic biological knowledge in terms of gene regulatory networks is taken into account. In this paragraph, instead, we restrict the analysis to the set of the 27 genes identified in the blood and tissue comparisons of Chapter 2 (see Table 2.3, last row) and the corresponding "high affinity" genes as found by performing random walks in the biological graph (Section 4.2.1). As described in Chapter 4 each of the 27 genes is allowed to grow a neighborhood of "close" enough genes (where the closeness is determined by the probability of transition based on the "random walks" in the input biological network) while the size of the neighborhoods is determined by the classification performance in a two level classification scheme. Each neighborhood is therefore determined based both on the biological characteristics and interactions of the genes, and on their classification and discriminating abilities on gene expression data sets. Additionally, the initial selection of the 27 seed genes is of course highly influential in the induced neighborhoods and we therefore expect that the seed genes and the corresponding neighborhoods have a great potential for producing biological relevant cluster assignments.

The extracted gene neighborhoods lend themselves quite naturally to the mixture of models setting that we are describing in this chapter. Effectively, each neighborhood delineates a set of genes that can be considered to be dependent and "cross-talking". Therefore, we consider each of the K "functional groups" of (5.3) to correspond to a single gene neighborhood. This means that genes in the same neighborhood are considered to be dependent whereas genes in distinct neighborhoods are believed to be independent, and these relationships are imposed by the structure of the covariance matrix in the Gaussian Mixture Model. In practice we end up with a few connections or common genes shared by neighborhoods. In order to address this issue, we try to break these connections among the neighborhoods by assigning these "shared" genes to the closest (minimum distance) neighborhood. The distance $D(i, j)$ of a gene i from the neighborhood N_j of a seed gene j is computed based on the sum of the transition probabilities from the neighbors of j to the i gene, as follows:

$$D(i, j) = -\log \sum_{k \in N_j, k \neq i} \Pr(k \to i) \tag{5.14}$$

where $\Pr(k \to i)$ is the random walk transition probability from gene k to gene i. This formula allows us to incorporate genes from the initial list of the 27 genes that were excluded from the construction of the neighborhoods because they were missing from the data set that we used to build the classifiers, and these are: SAR1A, CXCR4, SRSF6, EIF6. After these computations, we identify the totally separate gene neighborhoods shown in Figure 5.1 for the case of the CTC versus peripheral blood comparison and in Figure 5.2 for the case of the peripheral blood versus breast cancer tissue comparison. In the latter case, the neighborhoods contain much more genes but also all of the genes of the CTC versus peripheral blood comparison. For this reason, we decided to proceed using only the neighborhoods identified in the peripheral blood versus breast cancer tissue comparison (Figure 5.2).

For the evaluation of the network based Gaussian mixture model constructed by the gene neighborhoods as described above we use the GSE52604 dataset [Salhia et al., 2014] available from GEO[2] and the data set of Huang [Huang et al., 2003] that previously used in the previous paragraph. Both data sets refer to

[2] Available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52604 (accessed on May 25, 2016)

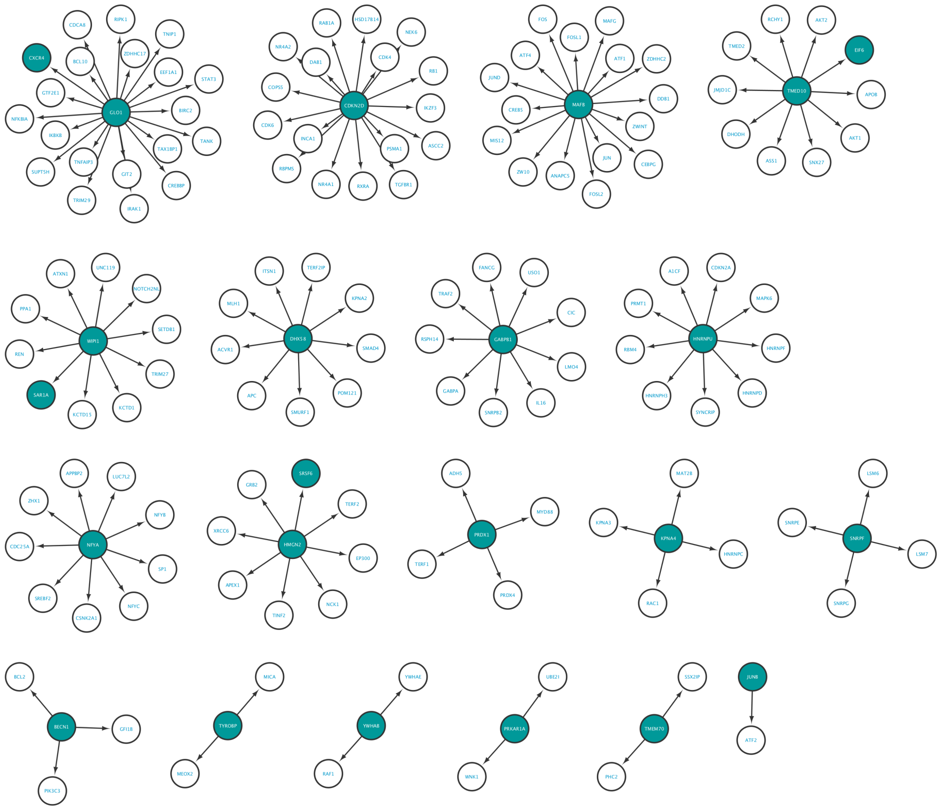


Figure 5.1: The neighborhoods of genes around the "seeding" genes after making them non-overlapping by assigning the common neighbors to the "closer" seed. These neighborhoods are the ones found in the CTC versus Peripheral Blood comparison (see Figure 4.3) with the addition of the four missing genes from the initial list of 27.

Breast Cancer patients and especially the GSE52604 contains 10 breast-brain metastatic samples and 10 non-neoplastic breast tissues.

For each evaluation data set we perform clustering using a "standard" "diagonal" Gaussian mixture model (GMM), where each Gaussian component considers independent genes but with different variances across the diagonal of the convariance matrix, and our "Stratified GMM" that considers the groups of genes according to the extracted neighborhoods. The block covariance matrix of (5.3) imposes a diagonal matrix for the genes not belonging in any group, but because the vast majority of the genes do not participate in the identified neighborhoods it seems plausible that the "Stratified GMM" provides similar results to the "Diagonal GMM". In order to reveal the potential differences in the two models we project the data sets to the list of genes participating in the neighborhoods of Figure 5.2, effectively performing a "crude" feature selection to the genes of interest. In order to further test the genes in the extracted neighborhoods we also create 1,000 random resamples on the original set of

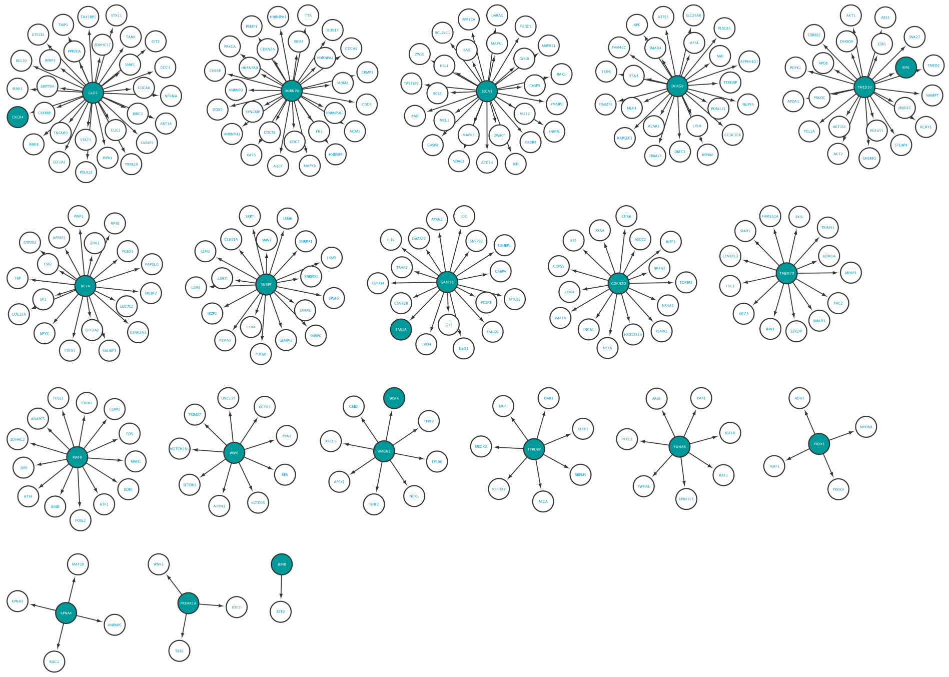


Figure 5.2: The neighborhoods of genes around the "seeding" genes after making them non-overlapping by assigning the common neighbors to the "closer" seed. These neighborhoods are the ones found in the Peripheral Blood versus Breast Cancer tissue (see Figure 4.4) with the addition of the four missing genes from the initial list of 27.

genes: each time a random subset of the genes with the same number of genes as in our neighborhoods is selected and then a Diagonal GMM is run using only this subset of features.

We are interested in the evaluation of the different mixture models with respect to their ability to produce clusters of samples with strong agreement with the biological characteristics of the said samples. For example, a perfect 2-cluster assignment of samples would allocate all healthy subjects in one cluster and all the non-healthy ones in the other. Nevertheless, the mixture models with the EM algorithm are easily trapped in a local maximum of the log likelihood function and their results can differ in each run due to the initialization of their parameters. The sensitivity to the initial parameter values is the reason that usually multiple runs or more advanced random initializations are suggested [McLachlan and Peel, 2000, Biernacki et al., 2003]. To overcome this issue, we give all algorithms the same initial parameters for the cluster means ("centroids") that are computed in supervised way based on the sample labels, as follows: we take half the samples of each class (for example, the metastatic samples) and estimate the class specific mean expression values based on these. Therefore, all mixture models start from the same means and the same sample covariance matrix.

Table 5.5: Cluster performance results in GSE52604

Mixture Model	BHI	loglik	AIC	BIC
Stratified	**1.0**	**-4,515.96**	21,433.91	32,637.02
Diagonal	0.90	-10,917.78	24,085.57	26,118.06
Random Resampling, Diagonal (mean ± std)	0.95 (±0.04)	-13,188.25 (±363.73)	28,626.50 (±727.46)	30,659.00 (±727.46)

For each mixture model tested we compute a number of metrics: the Biological Homogeneity Index (BHI) defined in Equation (5.13) above, the log likelihood of the parameters given the data, that is, the probability of the data set for the selected parameter values, and two standard criteria frequently used for model selection [Burnham and Anderson, 2004]: the Bayesian information criterion (BIC) and the Akaike information criterion (AIC). These indices are defined as follows:

$$\mathrm{BIC} = -2 \cdot \ln \mathcal{L}(\hat{\theta}) + p \cdot \ln(N) \tag{5.15}$$

$$\mathrm{AIC} = -2 \cdot \ln \mathcal{L}(\hat{\theta}) + 2 \cdot p \tag{5.16}$$

where $\mathcal{L}(\hat{\theta})$ is the value of the likelihood function on the estimated parameters θ, p is the number of the parameters of the model, and N is the sample size. For both information criteria smaller values represent better models, while they penalize models with high number of parameters with BIC taking into account the size of data set under consideration.

The results for GSE52606 are shown in Table 5.5. There is a clear indication that all models achieve almost excellent separation of the metastatic and non-metastatic samples in two clusters, as measured by the BHI metric. The Stratified model achieves excellent BHI, has the best (largest) likelihood of the data given its estimated parameters, and has the best (lower) AIC index value. Its BIC performance though is worse than the one of the Diagonal model due to the fact that it has a lot more parameters to estimate and the BIC index penalizes, more strongly that the AIC, the complexity of a given model (since it takes into account the sample size). Interestingly, the random models built during the resampling process achieve also very good performance, although the fit to the data and the AIC value is the worst in the set of candidate models.

We next proceed to the data set of [Huang et al., 2003] which relates to recurrence of cancer in breast cancer patients, with 18 samples of patients suffering a recurrence within three years after surgery and 34 samples without.

Table 5.6: Cluster performance results in Huang dataset

Mixture Model	BHI	loglik	AIC	BIC
Stratified	**0.74**	**-1,437.30**	12,332.61	21,560.04
Diagonal	0.60	-3,142.19	8,166.38	10,002.45
27 genes Diagonal	0.52	-365.34	916.68	1,098.15
27 genes Full	0.67	-237.83	1,673.65	2,842.45
Random Resampling, Diagonal (mean ± std)	0.63 (±0.04)	-2,427.42 (±334.61)	6,736.85 (±669.22)	8,572.97 (±669.22)

This appears to be a "difficult" data set and the very good classification results presented by the authors in the original publication could not be reproduced in a subsequent analysis [Ruschhaupt et al., 2004]. The results of the different clustering models are shown in Table 5.6. The Stratified mixture model takes the first place in terms of BHI, which, although not excellent, is still quite good. The Diagonal mixture model yields a good BHI but still no better than the average BHI of the random clustering models. Also its AIC and BIC performance is better than the Stratified model due to the number of parameters being almost 5 times less than in the Stratified model and the comparative values in the log likelihood function.

In Table 5.6 we also show the results when only the seeds responsible for the creation of the neighborhoods and the groups of genes that led to the Stratified model are used. Since we have only 27 genes we are able to test also the "full" model, that is, the one where a complete (non-sparse) covariance matrix is assumed, in addition to the diagonal case. The even less number of parameters to be estimated allows these models to have the best fit to the data and subsequently the best AIC and BIC scores. The BHI value, though, for the diagonal model is pretty average, but what is quite interesting is that when the full covariance matrix is assumed the BHI score is significantly increased.

Finally, the random resamples allows us to get a statistical view on the results of the Stratified model and the full GMM for the 27 genes. In Figure 5.3 we show a histogram of the random BHI values and the corresponding results of these models. The Full GMM using only the 27 genes yields a BHI score above the 80th percentile of the random values and a *p-value* of 0.123. The Stratified GMM achieves a score that only 4 out of the 1000 random samples match or exceed, for a p-value of 0.004.

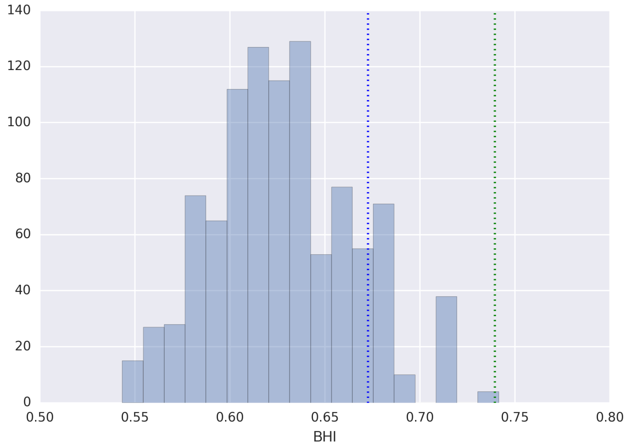


Figure 5.3: The distribution of the BHI values achieved by the Diagonal mixture model after random subsets of genes are selected 1000 times. The blue vertical line is the BHI value resulted from the full covariance mixture model for only the 27 genes, while the green vertical line is the corresponding value for the Stratified mixture model based on the extracted neighborhoods of these genes.

5.4 Conclusions and Future work

Early work on the analysis of gene expression data focused on the identification of a small number of informative genes whose expression levels are able to discriminate between different phenotypes or experimental conditions [Sotiriou and Pusztai, 2009]. In this chapter we have described a potential exploitation of existing biological information for the *integrated* analysis of gene expression data with the extracted 27 gene set and the induced gene “neighborhoods” used for evaluation. A justification of this approach is the recent advances in molecular biology that have suggested the study of the cell as a dynamic system, leading to the systems biology: instead of studying one gene or protein at a time, a more holistic approach is usually followed [Ideker et al., 2001].

The use of existing biological knowledge to guide data mining tasks in bioinformatics is definitely not a novel idea. Pan and colleagues have advocated it in a series of publications [Pan, 2006, Wei and Pan, 2008, Tai and Pan, 2007]. The structural model (5.3) of the covariance matrix has been independently studied by [Tritchler et al., 2009] while model-based approaches have also used elsewhere e.g. [Yeung et al., 2001]. Contrary to these efforts, the authors in [Jelizarow et al., 2010] argue quite convincingly that most of the publications in this line of thought show a considerable “over-optimism” [Jelizarow et al., 2010] and bias in their results by choosing their data sets, settings, or reporting only the “good” results. On this last point, we believe that at least we have been totally honest and open with the results of our approach.

In [Pan, 2006] the author considers K Gaussian mixture models, one for each block (category) of genes, and then proceeds to perform K EMs. The differences when compared to our approach are the following:

- Pan performs gene clustering instead of clustering the samples. Therefore he assumes a mixture model for the expression of genes where genes in the same group ("stratum") share the same prior probability coming from a given cluster.
- In our approach the clustering of samples takes place by considering group specific dependencies for the genes belonging to the same group, while the samples are considered i.i.d.[3] Instead, Pan assumes that the genes are independent but they share (gene) group specific prior probabilities for cluster membership.
- We have a more complex implementation due to the interdependences between the genes belonging in the same group, that is, we do not constrain the elements of each block sub-matrix in the covariance matrix shown in (5.3).

In terms of the assumptions of the proposed mixture model, the covariance model (5.3) assumes the independence of the uncategorized genes, an assumption that is definitely far from the truth. In fact, since the annotation of the human genome is an ongoing task our knowledge of the genes functionality and characteristics is pretty limited and changes every day. Nevertheless this is an simplifying assumption frequently encountered in gene expression analyses (e.g. Diagonal Linear Discriminant Analysis (LDA)) in order to deal with the high dimensionality and the danger of overfitting. The model (5.3) can also be seen as a middle solution between choosing the full sample covariance matrix, which can lead to an ill-posed inverse problem [Hansen, 1998], and a lower dimensional diagonal covariance matrix. On another note, it can be the case that certain genes can have more than one functional annotation or participate in more than one category or pathway. This is an additional observation that contradicts with the covariance structure of (5.3) since it means that there's inter group dependence. In this work, we were able to split the ties and create distinct groups of features based on external criteria (random walks probabilities) but more generic solutions may be sought in the future.

The presented approach has been tested in an unsupervised classification of human tumors in an exploratory stage of the analysis using the generic biological knowledge from a relatively large numbers of gene interaction pathways.

[3] Independent and Identically Distributed

The outcome of these experiments were not very informative on the validity of the described approach. The incorporation, though, of a more focused gene set and the genes strongly interacting with them (27 genes signature and their neighborhoods) presented a clear advantage of the methodology.

The "gene neighborhoods" of the 27 biomarkers identified in the previous chapter were the building blocks for the application of the mixture model that we present here, but in order to apply this model any shared genes among the neighbohoods are "attracted" to the "closer" one. We have also used the PANTHER[4] database for the post-assessment of the statistical overrepresentation of the resulted non-overlapping neighborhoods. Regarding the seed subnetworks as a whole entity in the case of CTC *versus* PB or PB *versus* BC comparison, we can clearly see various similarities and differences in terms of Gene Ontology (GO) biological processes and PANTHER pathways, which are governed the "neighborhood" genes of the base-classifiers (Tables 5.7 and 5.8). Since the "neighborhood" genes of the base classifiers in CTC *versus* PB comparison (160 genes) constitute a subset of the corresponding "neighborhood" genes in PB *versus* BC comparison (282 genes), the apparent similarity observed is expected and indirectly confirms the notion that CTCs carry information from the primary tumor, while any differences in CTC *versus* BC comparison could indicate a hidden metastatic potential of CTCs. Additionally, the overlapping processes and pathways in our CTC *versus* PB group expansion with the corresponding discriminant genes in the dataset of [Lang et al., 2015] unfold high similarities, further assessing the discriminative power on this expanded gene set.

Therefore, the resulting base-classifiers either individually or in combination (i) might be valuable for CTC tracking in the peripheral blood (ii) can shed light on the biological features and the molecular mechanisms of CTCs, and (ii) can provide operational models to test biological hypotheses underlying CTC status and metastatic potential.

[4] http://pantherdb.org/

Table 5.7: PANTHER Gene Ontology (GO) Biological Processes

	CTC *versus* PB		PB *versus* BC	
GO Slim Terms	Fold Enrichment	P-value	Fold Enrichment	P-value
RNA splicing, via transesterification reactions (GO:0000375)			8.45	1.31e-07
RNA splicing (GO:0008380)			8.26	1.78e-07
mRNA splicing, via spliceosome (GO:0000398)	5.72	2.06e-02	7.31	2.45e-08
mRNA processing (GO:0006397)	4.78	1.24e-02	6.51	1.99e-10
RNA metabolic process (GO:0016070)	2.55	3.06e-07	2.39	6.81e-11
nucleobase-containing compound metabolic process (GO:0006139)	2.11	2.61e-06	2.14	2.52e-12
primary metabolic process (GO:0044238)	1.57	1.71e-04	1.52	4.35e-07
metabolic process (GO:0008152)	1.43	2.47e-03	1.38	5.95e-05
response to stress (GO:0006950)	2.98	3.94e-02	3.61	1.28e-07
regulation of transcription from RNA polymerase II promoter (GO:0006357)	2.88	5.79e-05	2.14	1.98e-03
transcription from RNA polymerase II promoter (GO:0006366)	2.58	5.67e-05	2.20	1.94e-05
transcription, DNA-dependent (GO:0006351)	2.43	1.04e-04	2.07	5.48e-05
regulation of nucleobase-containing compound metabolic process (GO:0019219)	2.39	1.09e-03	1.88	1.09e-02
regulation of biological process (GO:0050789)	1.79	3.85e-02	1.65	9.17e-03
protein phosphorylation (GO:0006468)			2.96	6.55e-04
cellular protein modification process (GO:0006464)			1.92	5.00e-02
cell cycle (GO:0007049)	2.84	9.11e-04	2.95	3.37e-08
cellular process (GO:0009987)			1.40	1.22e-03
cell communication (GO:0007154)			1.61	1.36e-02
apoptotic process (GO:0006915)			2.58	4.27e-02

Table 5.8: PANTHER Pathways

Pathway	CTC *versus* PB		PB *versus* BC	
	Fold Enrichment	P-value	Fold Enrichment	P-value
Hypoxia response via HIF activation (P00030)	18.70	1.09e-02		
Toll receptor signaling pathway (P00054)	16.36	4.90e-05	13.27	1.10e-06
T cell activation (P00053)	16.36	1.37e-07	12.08	1.87e-08
TGF-beta signaling pathway (P00052)	15.32	3.84e-08	11.86	8.85e-10
Apoptosis signaling pathway (P00006)	13.66	2.10e-08	14.22	2.21e-16
p53 pathway (P00059)	12.93	4.07e-05	12.85	1.61e-09
Transcription regulation by bZIP transcription factor (P00055)	12.83	7.90e-03	14.58	4.57e-07
B cell activation (P00010)	12.08	1.94e-03	10.29	5.21e-05
Ras Pathway (P04393)	11.60	4.71e-04	10.35	2.43e-06
Interleukin signaling pathway (P00036)	10.80	1.55e-04	8.43	1.88e-05
CCKR signaling map (P06959)	10.07	1.37e-07	10.56	4.43e-15
PDGF signaling pathway (P00047)	9.49	2.20e-05	8.08	1.66e-07
Gonadotropin releasing hormone receptor pathway (P06664)	9.31	4.55e-09	7.60	1.88e-11
EGF receptor signaling pathway (P00018)	8.06	1.31e-03	6.86	4.43e-05
Angiogenesis (P00005)	7.65	5.37e-04	6.76	5.43e-06
FGF signaling pathway (P00021)	7.45	7.90e-03	6.65	1.88e-04
Inflammation mediated by chemokine and cytokine signaling pathway (P00031)	5.88	5.42e-04	4.25	1.23e-03
p53 pathway feedback loops 2 (P04398)			11.66	9.39e-05
Insulin/IGF pathway-protein kinase B signaling cascade (P00033)			11.44	2.73e-03
VEGF signaling pathway (P00056)			10.08	2.75e-04
Endothelin signaling pathway (P00019)			6.84	4.45e-03

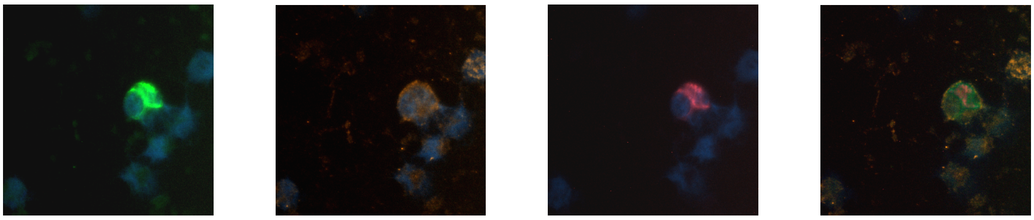
Chapter 6

Conclusions

In this thesis we focus on the identification of novel marker genes that provide insights on the differentiating characteristics of tissue and peripheral blood samples of breast cancer patients. The underlying biological justification of this differentiation is the circulating tumor cells in the peripheral blood of the patients. These circulating tumor cells have long been correlated with the relapse and triggering of the metastatic cascade of the disease. The approach we followed was based on the computational and statistical methodologies for the analysis of gene expression data produced by DNA microarrays. Our work was based therefore more on a data-driven, exploratory methodology rather than on the biology-driven scientific experiments performed by the domain experts (molecular biologists and biochemists).

As part of this effort, we have arrived in a limited set of potential biomarkers (genes) that exhibit elevated expression in cancer peripheral blood. These findings were the result of the pooling of a large number of gene expression data from different studies that were homogenized and integrated in order to be put over a common statistical foundation. This was a time consuming and laborious task that required a lot of diligence in order to overcome a number of technical and domain specific challenges. The statistical methodology was largely based on the comparative assessment of the expression of the genes into different conditions (control and cancerous samples, tissue and peripheral blood sites) while controlling, at the same time, for possible biases introduced by the separate experimental conditions. We consider the methodology and its results to be the first main contributions of this thesis. A first evaluation of the 27 potential biomarkers using existing biological knowledge such the GO and KEGG databases provided encouraging results.

The following step was to expand the list of the 27 genes using the information encoded in the biological networks. Using graph algorithms we succeeded in expanding the initial list of genes exploiting both local properties of the



(a) Cytokeratin *(b) CXCR4* *(c) JUNB* *(d) Overlay*

Figure 6.1: Expression of Cytokeratin (CK), CXCR4 and JUNB in CTCs isolated from breast cancer patients.

network and probabilistic "random walks" in it. The result is an enriched set of genes with potentially differentiating activity that combine the knowledge coming from the biology with the discriminating power of our initial "seed" list of genes. The use of this augmented knowledge is then assessed into two common tasks in machine learning and data mining: supervised and unsupervised learning. The second major area of contributions is therefore the use of prior knowledge in order to increase the performance of these tasks.

In this research effort we have addressed a number of computational and domain-specific issues, such as the diversity of the samples in the public data sets, the heterogeneity of the microarray platforms, the different annotations, and so on, in order to arrive to a limited and statistically significant set of genes. Of course, it is common knowledge that statistical significance of any findings do not always mean significant effect sizes. Nevertheless, a targeted biological experiment performed in the Cancer Biology Laboratory, School of Medicine, University of Crete, focusing on a specific pathway containing a subset of the 27 genes, yielded results that are in agreement to our bioinformatics analysis. In particular, two of our 27 genes, **CXCR4**, a chemokine receptor which is involved in tumor metastasis and **JUNB** a transcription factor participating in CXCR4 pathway[1], were evaluated in samples from metastatic breast cancer patients, cell lines, and CTCs [Kallergi et al., 2015].

In more detail, triple staining immunofluoresence with panytokeratin/CXCR4/JUNB antibodies were performed in SKBR3, MDA-MB231, MCF7 and Hela cell lines. The same experiments were performed in Peripheral Blood Mononuclear Cells (PBMCs) from normal ($n = 10$) subjects and in PBMCs ($n = 55$) from untreated metastatic breast cancer patients. Statistical analysis revealed significant differences in the expression of both molecules between healthy donors' PBMCs and patient's PBMCs. Subsequently, CTCs were detected in 17 out of 55 screened patients (Figure 6.1). Patients with CXCR4-positive CTCs (with mean expression higher than 95% of normal

[1] This pathway includes: JUNB $\rightleftharpoons$ BRCA1 $\rightleftharpoons$ JAK2 $\rightleftharpoons$ CXCR4

PBMCs) were 53%. In addition 84% of the examined CTCs had expression profile higher than normal PBMCs. Likewise, JUNB expression in CTCs (above 95% of normal PBMCs) was identified in 76.92% of patients. Furthermore 64.3% of the total CTCs have expression higher than Normal PBMCs. Therefore, the results of this experiment show that CXCR4 and JUNB are highly expressed in CTCs derived from breast cancer patients, in agreement to bioinformatics' analysis. Quantification of immunofluoresence potentially delineates a subgroup of patients with high expression of CXCR4 and JUNB that could benefit from target therapies.

In conclusion, further validation of the results is always possible, which can also open new avenues for research. New data sets, new technological developments, and theoretical and technical advancements continuously emerge. We believe that the methodologies described in this dissertation are generic and there is high probability that they can be adapted in future requirements.

Appendix A

SVD and Reduced Rank approximation

We have an $N x M$ real data matrix X where features (genes) are in the columns and the cases (samples, DNA arrays) are in the rows. We extract the column means from each row to make it "centered":

$$\widetilde{X} = X - \mathbf{1}_N \boldsymbol{\mu}^T \tag{A.1}$$

and then the (sample) covariance matrix is given by

$$\Sigma \equiv \mathbb{E}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T = \frac{1}{N-1}\widetilde{X}^T\widetilde{X} \tag{A.2}$$

We make use of the Singular Value Decomposition (SVD) of the data matrix [Meyer, 2000]:

$$\widetilde{X} = UDV^T \tag{A.3}$$

where U and V are $N x M$ and $M x M$ orthogonal matrices (i.e. $UU^T = \mathrm{I}$ and $VV^T = \mathrm{I}$) and D is an $M x M$ diagonal matrix with its diagonal elements being the "singular values" of X with $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_r > 0$ where r is the rank of matrix $\widetilde{X}$. Now based on the SVD of $\widetilde{X}$, the covariance matrix becomes

$$\Sigma = \frac{1}{N-1} VDU^TUDV^T = \frac{1}{N-1} VD^2V^T \tag{A.4}$$

Therefore the $M x r$ matrix V contains, in its columns, the r orthonormal eigenvectors of the covariance matrix while its eigenvalues are given by

$$\lambda_i = \frac{1}{N-1}\sigma_i^2 \tag{A.5}$$

while the "total variance" in the data is:

$$\mathrm{tr}(\Sigma) = \frac{1}{N-1}\mathrm{tr}(VD^2V^T) = \frac{1}{N-1}\mathrm{tr}(D^2VV^T) = \frac{1}{N-1}\sum_{k=1}^{r}\sigma_k^2 \tag{A.6}$$

Now as a dimension reduction technique we can use the transformation:

$$y = V^T \tilde{x} \tag{A.7}$$

so that a sample x, after centering, is mapped from the $\mathbb{R}^M$ space to the smaller $\mathbb{R}^r$. This transformation in matrix form is given by (remember that the X has the cases as rows while $\tilde{x}$ above is a column vector for a single case):

$$Y = \tilde{X}V \tag{A.8}$$

and has covariance matrix

$$\Sigma_Y = \frac{1}{N-1}Y^TY = \frac{1}{N-1}V^T\tilde{X}^T\tilde{X}V = \frac{1}{N-1}V^TVD^2V^TV = \frac{1}{N-1}D^2 \tag{A.9}$$

i.e. it is diagonal with diagonal elements $\frac{1}{N-1}\sigma_i^2$ and of r rank ($\sigma_i = 0, i > r$).

The transformation A.8 is actually the Principal Component Analysis technique where we are using all the eigenvalues of the covariance matrix. If, instead, we keep the first K eigenvectors that correspond to the largest eigenvalues (or, equivalently, the largest squares of singular values per A.5), the transformed centred data will be:

$$Y_K = \tilde{X}V_K \tag{A.10}$$

where V_K has the K eigenvectors in its columns. The corresponding covariance matrix uses the first K singular values:

$$\Sigma_{Y_K} = \frac{1}{N-1}D_K^2 \tag{A.11}$$

The "error" of this "low rank" transformation in terms of the Frobenius form would be (matrix approximation lemma [Eckart and Young, 1936]):

$$||Y - Y_K||_F^2 = \frac{1}{N-1}\sum_{k=K+1}^{r} \sigma_k^2 \tag{A.12}$$

A.1 Use in rank-deficient Gaussian distributions

The probability density function (pdf) of a multivariate (M-dimensional) normal distribution is given by:

$$f(\mathbf{x}) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{A.13}$$

In the analysis of gene expression and other high throughput data, it is always the case that the number of samples is lower than the number of features (e.g. genes), i.e. $N \ll M$. In these cases therefore, the covariance matrix Σ is not full rank, and the above formula can not be used. In order to bypass this problem we can use the SVD of the original data matrix as described above.

Concretely, the following matrix is the Moore-Penrose pseudoinverse of the covariance matrix defined in Equation A.4:

$$\Sigma^+ = (N-1)VD^{-2}V^T \tag{A.14}$$

This matrix is built using the r non-zero entries of the diagonal matrix D (Equation A.3) and satisfies the properties of the pseudoinverse (e.g. $\Sigma\Sigma^+\Sigma = \Sigma$, $\Sigma^+\Sigma\Sigma^+ = \Sigma^+$ etc.) [Golub and Van Loan, 2012]. Using this in place of the original convariance matrix in Equation 5.4 (and its determinant $|\Sigma^+| = (N-1)^r \prod_{i=1}^{r} \sigma_i^{-2}$) allows the computation of the Gaussian pdf.

Appendix B

The "Block diagonal" EM algorithm

In the statified model we assume that genes can be classified in K different categories using GO Biological Processes for example. We make the assumptions that genes in different categories are independent and the covariance matrix (by using some proper rearrangement of the genes[1]) is a block diagonal matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}^{(1)} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Sigma}^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\Sigma}^{(K)} \end{bmatrix} \tag{B.1}$$

If we modelled the data by a single Gaussian

$$f_i(\mathbf{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{L/2}|\boldsymbol{\Sigma}_i|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_j - \boldsymbol{\mu}_i)} \tag{B.2}$$

then taking into account the block triagonal structure of (B.1) the normal density of (B.2) factorizes into

$$f_i(\mathbf{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \prod_{k=1}^{K} \frac{1}{(2\pi)^{N_k/2}|\boldsymbol{\Sigma}_i^{(k)}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)})^T \boldsymbol{\Sigma}_i^{(k),-1}(\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)})} \tag{B.3}$$

where we have used the "exponent" (k) to refer to the projection/selection of the genes (and means, covariance matrices) belonging to the k category.

Now we take a mixture of Gaussians like the (B.3) the equation (5.1) becomes

$$f(\mathbf{x}_j; \Theta) = \sum_{i=1}^{g} \pi_i \prod_{k=1}^{K} \frac{1}{(2\pi)^{N_k/2}|\boldsymbol{\Sigma}_i^{(k)}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)})^T \boldsymbol{\Sigma}_i^{(k),-1}(\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)})} \tag{B.4}$$

[1] ..and possibly repeating a gene if it's classified in more than one category...

that can be written more compactly as

$$f(\mathbf{x}_j;\Theta) = \sum_{i=1}^{g} \pi_i \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)}) \tag{B.5}$$

Now as in Section 5.2.1 we assume the “hidden” (missing) data $\mathbf{z}_j$ where $z_{ji} = 1$ if the j-th sample is generated by the i-th cluster and we can model the probability of the data as

$$\begin{aligned} f(\mathbf{x}_j, \mathbf{z}_j|\Theta) &= f(\mathbf{x}_j|\mathbf{z}_j,\Theta)f(\mathbf{z}_j|\Theta) \\ &= \prod_{i=1}^{K} (f_i(\mathbf{x}_j|\mathbf{z}_j,\Theta))^{z_{ik}} \prod_{i=1}^{K} (f_i(\mathbf{z}_j|\theta))^{z_{ik}} \\ &= \prod_{i=1}^{K} (f_i(\mathbf{x}_j|\mathbf{z}_j,\Theta)p_i(\mathbf{z}_j|\Theta))^{z_{ik}} \\ &= \prod_{i=1}^{K} (\pi_i f_i(\mathbf{x}_j|\mathbf{z}_j,\Theta))^{z_{ik}} \end{aligned} \tag{B.6}$$

(where π_i is the probability of cluster i, and the products appear since all but one of the z_{ik} will be zero.)

The complete data log likelihood based on (B.6) then becomes

$$\begin{aligned} \ln p(X, Z; \Theta) &= \ln \prod_{j=1}^{N} f(\mathbf{x}_j, \mathbf{z}_j|\Theta) \\ &= \ln \prod_{j=1}^{N} \prod_{i=1}^{g} (\pi_i f_i(\mathbf{x}_j|\mathbf{z}_j,\Theta))^{z_{ji}} \\ &= \sum_{j=1}^{N} \sum_{i=1}^{g} z_{ji} (\ln \pi_i + \ln f_i(\mathbf{x}_j|\mathbf{z}_j,\Theta)) \\ &= \sum_{j=1}^{N} \sum_{i=1}^{g} z_{ji} \left(\ln \pi_i + \sum_{k=1}^{K} \ln \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)}) \right) \end{aligned}$$

So the expectation to be maximized is

$$
\begin{aligned}
\mathcal{Q}(\Theta, \Theta^{\text{cur}}) &= \sum_{z} P(Z|X; \Theta^{\text{cur}}) \ln p(X, Z; \Theta) \\
&= \sum_{z} P(Z|X; \Theta^{\text{cur}}) \sum_{j=1}^{N} \sum_{i=1}^{g} z_{ji} \left(\ln \pi_i + \sum_{k=1}^{K} \ln \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)}) \right) \\
&= \sum_{j=1}^{N} \sum_{i=1}^{g} P(z_{ji} = 1 | \mathbf{x}_j; \Theta^{\text{cur}}) \left(\ln \pi_i + \sum_{k=1}^{K} \ln \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)}) \right) \\
&= \sum_{j=1}^{N} \sum_{i=1}^{g} \frac{p(\mathbf{x}_j | z_{ji} = 1; \Theta^{\text{cur}}) P(z_{ji} = 1; \Theta^{\text{cur}})}{\sum_{s=1}^{g} p(\mathbf{x}_j | z_{js} = 1; \Theta^{\text{cur}}) P(z_{js} = 1; \Theta^{\text{cur}})} \left(\ln \pi_i + \sum_{k=1}^{K} \ln \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)}) \right. \\
&= \sum_{j=1}^{N} \sum_{i=1}^{g} \frac{f_i(\mathbf{x}_j; \boldsymbol{\mu}_i^{\text{cur}}, \boldsymbol{\Sigma}_i^{\text{cur}}) \pi_i^{\text{cur}}}{\sum_{s=1}^{g} f_s(\mathbf{x}_j; \boldsymbol{\mu}_s^{\text{cur}}, \boldsymbol{\Sigma}_s^{\text{cur}}) \pi_s^{\text{cur}}} \left(\ln \pi_i + \sum_{k=1}^{K} \ln \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)}) \right) \\
&= \sum_{j=1}^{N} \sum_{i=1}^{g} \underbrace{\frac{\pi_i^{\text{cur}} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_i^{(k),\text{cur}}, \boldsymbol{\Sigma}_i^{(k),\text{cur}})}{\sum_{s=1}^{g} \pi_s^{\text{cur}} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_s^{(k),\text{cur}}, \boldsymbol{\Sigma}_s^{(k),\text{cur}})}}_{\tau_{ji}} \left(\ln \pi_i + \sum_{k=1}^{K} \ln \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)} \right. \\
&= \sum_{j=1}^{N} \sum_{i=1}^{g} \tau_{ji} \left(\ln \pi_i + \sum_{k=1}^{K} \ln \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)}) \right) \\
&= \sum_{j=1}^{N} \sum_{i=1}^{g} \tau_{ji} \ln \left(\pi_i \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)}) \right)
\end{aligned}
$$

where τ_{ji} is the (current) estimate of the probability the j sample was generated by (or belongs to) the i cluster

$$
\tau_{ji} = \frac{\pi_i^{\text{cur}} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_i^{(k),\text{cur}}, \boldsymbol{\Sigma}_i^{(k),\text{cur}})}{\sum_{s=1}^{g} \pi_s^{\text{cur}} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_j^{(k)}; \boldsymbol{\mu}_s^{(k),\text{cur}}, \boldsymbol{\Sigma}_s^{(k),\text{cur}})} \tag{B.7}
$$

For the M-step we need to find the parameters $\Theta = \{\pi_i, \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)}\}$, $i = 1 \ldots g$, $k = 1 \ldots K$ that maximize (B) under the constraint that $\sum_i \pi_i = 1$. We insert a Langrangian multiplier λ and try to maximize the expression

$$
J(\Theta, \Theta^{\text{cur}}) = \mathcal{Q}(\Theta, \Theta^{\text{cur}}) + \lambda \left(\sum_{i} \pi_i - 1 \right) \tag{B.8}
$$

We take the partial derivatives (see Section B.1 for some useful identities used) and set the result to zero

$$\frac{\partial J(\Theta,\Theta^{\text{cur}})}{\partial \boldsymbol{\mu}_i^{(k)}} = \sum_{j=1}^{N} \frac{\tau_{ji}}{\pi_i \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_j^{(k)};\boldsymbol{\mu}_i^{(k)},\boldsymbol{\Sigma}_i^{(k)})} \left(\frac{\partial \pi_i \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_j^{(k)};\boldsymbol{\mu}_i^{(k)},\boldsymbol{\Sigma}_i^{(k)})}{\partial \boldsymbol{\mu}_i^{(k)}} \right)$$

$$= \sum_{j=1}^{N} \tau_{ji} \left(-\frac{1}{2}\right) \frac{\partial}{\partial \boldsymbol{\mu}_i^{(k)}} (\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)})^T \boldsymbol{\Sigma}_i^{(k),-1} (\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)})$$

$$= \sum_{j=1}^{N} \tau_{ji} \left(-\frac{1}{2}\right) 2\boldsymbol{\Sigma}_i^{(k),-1} (\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)})$$

$$= \sum_{j=1}^{N} \tau_{ji} \boldsymbol{\Sigma}_i^{(k),-1} (\boldsymbol{\mu}_i^{(k)} - \mathbf{x}_j^{(k)})$$

Setting that to zero and multiplying by $\boldsymbol{\Sigma}^{(k)}$ (we assume that $\boldsymbol{\Sigma}^{(k)}$ is non-singular) we get

$$\boldsymbol{\mu}_i^{(k)} = \frac{\sum_{j=1}^{N} \tau_{ji} \mathbf{x}_j^{(k)}}{\sum_{j=1}^{N} \tau_{ji}} \tag{B.9}$$

Similarly, for the covariance matrices:

$$\frac{\partial J(\Theta,\Theta^{\text{cur}})}{\partial \boldsymbol{\Sigma}_i^{(k)}} = \sum_{j=1}^{N} \tau_{ji} \frac{\partial \ln \left(\pi_i \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_j^{(k)};\boldsymbol{\mu}_i^{(k)},\boldsymbol{\Sigma}_i^{(k)})\right)}{\partial \boldsymbol{\Sigma}_i^{(k)}}$$

$$= \sum_{j=1}^{N} \tau_{ji} \frac{\partial}{\partial \boldsymbol{\Sigma}_i^{(k)}} \left(\ln \mathcal{N}(\mathbf{x}_j^{(k)};\boldsymbol{\mu}_i^{(k)},\boldsymbol{\Sigma}_i^{(k)}) \right)$$

$$= \sum_{j=1}^{N} \tau_{ji} \frac{\partial}{\partial \boldsymbol{\Sigma}_i^{(k)}} \left(-\frac{N_k}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i^{(k)}| - \frac{1}{2} (\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)})^T \boldsymbol{\Sigma}_i^{(k),-1} (\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)}) \right)$$

$$= \sum_{j=1}^{N} \tau_{ji} \left(-\frac{1}{2} (\boldsymbol{\Sigma}_i^{(k),-1})^T - \frac{1}{2} \frac{\partial}{\partial \boldsymbol{\Sigma}_i^{(k)}} (\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)})^T \boldsymbol{\Sigma}_i^{(k),-1} (\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)}) \right)$$

$$= \sum_{j=1}^{N} \tau_{ji} \left(-\frac{1}{2} \boldsymbol{\Sigma}_i^{(k),-1} + \frac{1}{2} \boldsymbol{\Sigma}_i^{(k),-1} (\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)}) (\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)})^T \boldsymbol{\Sigma}_i^{(k),-1} \right)$$

and again setting that to zero and multiplying by $\boldsymbol{\Sigma}_i^{(k)}$ we get

$$\boldsymbol{\Sigma}_i^{(k)} = \frac{\sum_{j=1}^{N} \tau_{ji} (\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)}) (\mathbf{x}_j^{(k)} - \boldsymbol{\mu}_i^{(k)})^T}{\sum_{j=1}^{N} \tau_{ji}} \tag{B.10}$$

Finally, for the “mixing coefficients” we have:

$$
\begin{aligned}
\frac{\partial J(\Theta, \Theta^{\mathrm{cur}})}{\partial \pi_i} &= \sum_{j=1}^{N} \tau_{ji} \frac{1}{\pi_i} + \lambda = 0 \\
\Longrightarrow \lambda \pi_i &= -\sum_{j=1}^{N} \tau_{ji} \\
\Longrightarrow \lambda \sum_i \pi_i &= -\sum_{j=1}^{N} \sum_i \tau_{ji} \\
\Longrightarrow \lambda &= -N
\end{aligned}
$$

and therefore

$$
\pi_i = \frac{\sum_{j=1}^{N} \tau_{ji}}{N} \tag{B.11}
$$

In conclusion, using the stratified model as defined in the beginning of this section (with block diagonal covariance matrix) we end up with ML estimations of $\boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)}$ with the same structure (e.g. the mean $\boldsymbol{\mu}_i^{(k)}$ of mixture component i in the category k is computed by considering only the genes in this category $\mathbf{x}^{(k)}$)

B.1 Matrix Calculus

In the above calculations we have used the following matrix identities:

$$
\begin{aligned}
\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} &= (\mathbf{A} + \mathbf{A}^T)\mathbf{x} \\
\frac{\partial \mathbf{a}^T \mathbf{X}^{-1} \mathbf{b}}{\partial \mathbf{X}} &= -\mathbf{X}^{-T} \mathbf{a} \mathbf{b}^T \mathbf{X}^{-T} \\
\frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} &= |\mathbf{X}| (\mathbf{X}^{-1})^T \\
\frac{\partial \ln |\mathbf{X}|}{\partial \mathbf{X}} &= (\mathbf{X}^{-1})^T
\end{aligned}
$$

Appendix C

Publications

Contents

The research we present in this document was largely documented in various scientific publications. Here, we provide details on these publications.

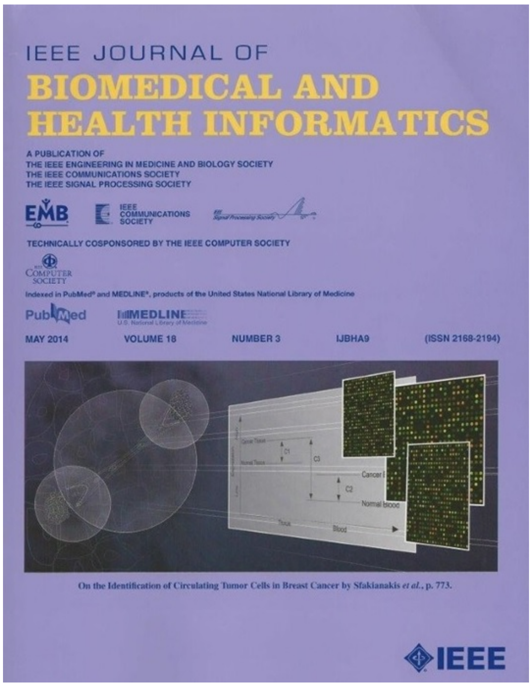
C.1 Journal papers

In the course of my research the following papers were included in scientific journals:

- [Sfakianakis et al., 2010a] is a review paper, providing an overview of the domain and the problems in the analysis of high throughput, gene expression data (In *International Journal of Biomedical Engineering and Technology*)

- [Sfakianakis et al., 2014] was the major output of the thesis presenting our effort and results for characterising the gene expression profile of the Circulating Tumor Cells using bioinformatics methods (In *IEEE Journal of Biomedical and Health Informatics*). The paper was also selected to appear on the cover page of the journal (Figure C.1). The contents of this paper constitute the backbone of Chapter 2.

- [Kallergi et al., 2015] presented the experimental validation of a subset of the genes identified in [Sfakianakis et al., 2014] (In *Cancer Research*).



IEEE JOURNAL OF
BIOMEDICAL AND
HEALTH INFORMATICS

A PUBLICATION OF
THE IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY
THE IEEE COMMUNICATIONS SOCIETY
THE IEEE SIGNAL PROCESSING SOCIETY

TECHNICALLY COSPONSORED BY THE IEEE COMPUTER SOCIETY

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

MAY 2014 VOLUME 18 NUMBER 3 IJBHA9 (ISSN 2168-2194)

On the Identification of Circulating Tumor Cells in Breast Cancer by Sfakianakis *et al.*, p. 773.

Figure C.1: The cover page of the IEEE Journal of Biomedical and Health Informatics, May 2014, featuring the [Sfakianakis et al., 2014] paper.

- [Sfakianakis et al., 2016a] is under preparation as an extended version of [Sfakianakis et al., 2016b].

Finally, [Notas et al., 2015] is a paper where I have contributed on the part related to the bioinformatics and statistical analysis (In *Molecular Oncology*). Although, not related to the core of this thesis, my contribution was largely based on the experience gained during the course of my research.

C.2 Papers in conference proceedings

The following thesis-related papers have been presented in conferences and are included in the corresponding proceedings:

- [Sfakianakis et al., 2010b] was a preliminary work for model based clustering using the mixture of Gaussians as the underlying model in a biological information driven way. More specifically, we were using the information from biological networks to impose a sparse solution using a "stratified" (or block-ed) version of the Expectation Maximization (EM) algorithm. Chapter 5 is largely based on this publication.

- [Sfakianakis et al., 2015] presented the "Steiner tree" expansion of the genes identified in [Sfakianakis et al., 2014] and it's expanded in Chapter 3.

- [Sfakianakis et al., 2016b] built upon the same list of initial "seed" genes and using the background network information constructed a two level classification scheme using adaptive, data-driven learning. Furthermore, the "neighborhoods" of the seed genes are used for *model-based clustering* with the modified "Expectation Maximization" fitting algorithm of [Sfakianakis et al., 2010b]. This work is described in detail in Chapter 4.

The compendium of the gene expression data sets gathered in this work has been used extensively. The following is a list of relevant publications that were based on this extended data set and address research questions related to the work presented in this thesis:

- F. Gypas, E. S. Bei, M. Zervakis and S. Sfakianakis, "A disease annotation study of gene signatures in a breast cancer microarray dataset," Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, Boston, MA, 2011, pp. 5551-5554. [Gypas et al., 2011]

- N. K. Chlis, S. Sfakianakis, E. S. Bei and M. Zervakis, "A generic framework for the elicitation of stable and reliable gene expression signatures," Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on, Chania, 2013, pp. 1-6. [Chlis et al., 2013]

- N. K. Chlis, S. Sfakianakis, E. S. Bei, D. Iliopoulou, D. Kafetzopoulos, and M. Zervakis, "Searching for Significant Genes in Cancer Metastasis by Tissue Comparisons." In 6th European Conference of the International Federation for Medical and Biological Engineering, pp. 594-597. Springer International Publishing, 2015. [Chlis et al., 2015]

- S. Tsakaneli, E. S. Bei, and M. Zervakis. "Comparing genomic network methodologies: A combined approach for cancer prognosis." In XIV Mediterranean Conference on Medical and Biological Engineering and Computing, pp 506-511. Springer International Publishing, 2016. [Tsakaneli et al., 2016]

- A. Alevyzaki, S. Sfakianakis, E. S. Bei, E. Obermayr, R. Zeillinger, D. Fotiadis, and M. Zervakis. "Biclustering strategies for genetic marker selection in gynecologic tumor cell lines." In 38th Annual International

Conference of the IEEE Engineering in Medicine and Biology Society, 2016. [Alevyzaki et al., 2016]

References

[Aaroe et al., 2010] Aaroe, J., Lindahl, T., Dumeaux, V., Saebo, S., Tobin, D., Hagen, N., Skaane, P., Lonneborg, A., Sharma, P., and Borresen-Dale, A.-L. (2010). Gene expression profiling of peripheral blood cells for early detection of breast cancer. *Breast Cancer Res*, 12(1):R7. (Cited on page 47.)

[Abraham et al., 2010] Abraham, G., Kowalczyk, A., Loi, S., Haviv, I., and Zobel, J. (2010). Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. *BMC bioinformatics*, 11(1):1. (Cited on page 37.)

[Akey et al., 2007] Akey, J. M., Biswas, S., Leek, J. T., and Storey, J. D. (2007). On the design and analysis of gene expression studies in human populations. *Nature Genetics*, 39(7):807–808. (Cited on page 29.)

[Alevyzaki et al., 2016] Alevyzaki, A., Sfakianakis, S., Bei, E. S., Obermayr, E., Zeillinger, R., Fotiadis, D., and Zervakis, M. (2016). Biclustering strategies for genetic marker selection in gynecologic tumor cell lines. In *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1–4. (Cited on pages 21 and 130.)

[Allard et al., 2004] Allard, W. J., Matera, J., Miller, M. C., Repollet, M., Connelly, M. C., Rao, C., Tibbe, A. G., Uhr, J. W., and Terstappen, L. W. (2004). Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases. *Clinical Cancer Research*, 10(20):6897–6904. (Cited on page 13.)

[Alliance, 2010] Alliance, G. (2010). Understanding genetics: A district of columbia guide for patients and health professionals. *Washington (DC): Genetic Alliance*. (Cited on page 4.)

[Alm and Arkin, 2003] Alm, E. and Arkin, A. P. (2003). Biological networks. *Current opinion in structural biology*, 13(2):193–202. (Cited on page 32.)

[Alon, 2003] Alon, U. (2003). Biological networks: the tinkerer as an engineer. *Science*, 301(5641):1866–1867. (Cited on page 32.)

[Anderson et al., 2010] Anderson, W. F., Jatoi, I., Tse, J., and Rosenberg, P. S. (2010). Male breast cancer: a population-based comparison with female breast cancer. *Journal of Clinical Oncology*, 28(2):232–239. (Cited on page 6.)

[Ashburner et al., 2000] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25. (Cited on pages 32 and 94.)

[Ashworth, 1869] Ashworth, T. R. (1869). A case of cancer in which cells similar to those in the tumours seen in the blood after death. *Australian Medical Journal*, 14:146–149. (Cited on pages 9 and 10.)

[Azuaje and Bodenreider, 2004] Azuaje, F. and Bodenreider, O. (2004). Incorporating ontology-driven similarity knowledge into functional genomics: An exploratory study. In *Bioinformatics and Bioengineering, 2004. BIBE 2004. Proceedings. Fourth IEEE Symposium on*, pages 317–324. IEEE. (Cited on page 32.)

[Baldi and Brunak, 2001] Baldi, P. and Brunak, S. (2001). *Bioinformatics: the machine learning approach.* MIT press. (Cited on page 3.)

[Balmain et al., 2003] Balmain, A., Gray, J., and Ponder, B. (2003). The genetics and genomics of cancer. *Nature Genetics*, 33(3s):238–244. (Cited on page 46.)

[Banerjee and El Ghaoui, 2008] Banerjee, O. and El Ghaoui, L. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516. (Cited on page 96.)

[Banfield and Raftery, 1993] Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821. (Cited on page 20.)

[Barabási et al., 2011] Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68. (Cited on pages 33 and 74.)

[Barbazán et al., 2012] Barbazán, J., Alonso-Alconada, L., Muinelo-Romay, L., Vieito, M., Abalo, A., Alonso-Nocelo, M., Candamio, S., Gallardo, E., Fernández, B., Abdulkader, I., de Los Ángeles Casares, M., Gómez-Tato, A., López-López, R., and Abal, M. (2012). Molecular characterization of circulating tumor cells in human metastatic colorectal cancer. *PloS one*, 7(7):e40476. (Cited on pages 36, 45 and 46.)

[Bard and Rhee, 2004] Bard, J. B. and Rhee, S. Y. (2004). Ontologies in biology: design, applications and future challenges. *Nature Reviews Genetics*, 5(3):213–222. (Cited on page 32.)

[Barrett et al., 2010] Barrett, K. E., Barman, S. M., Boitano, S., et al. (2010). *Ganong's review of medical physiology.* New Delhi: McGraw Hill, 2010. (Cited on page 35.)

[Barrett et al., 2005] Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W., and Edgar, R. (2005). NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic acids research*, 33(Database issue):D562–6. (Cited on page 56.)

[Benjamini and Hochberg, 1995a] Benjamini, Y. and Hochberg, Y. (1995a). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B ...*. (Cited on pages 19, 52 and 68.)

[Benjamini and Hochberg, 1995b] Benjamini, Y. and Hochberg, Y. (1995b). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B ...*. (Cited on page 49.)

[Bettegowda et al., 2014] Bettegowda, C., Sausen, M., Leary, R. J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B. R., Wang, H., Luber, B., Alani, R. M., et al. (2014). Detection of circulating tumor DNA in early-and late-stage human malignancies. *Science translational medicine*, 6(224):224ra24–224ra24. (Cited on page 37.)

[Beyer et al., 1999] Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? In *Database theory—ICDT'99*, pages 217–235. Springer. (Cited on page 24.)

[Bidard et al., 2014] Bidard, F.-C., Peeters, D. J., Fehm, T., Nolé, F., Gisbert-Criado, R., Mavroudis, D., Grisanti, S., Generali, D., Garcia-Saenz, J. A., Stebbing, J., et al. (2014). Clinical validity of circulating tumour cells in patients with metastatic breast cancer: a pooled analysis of individual patient data. *The Lancet Oncology*, 15(4):406–414. (Cited on page 11.)

[Bidard et al., 2013] Bidard, F.-C., Pierga, J.-Y., Soria, J.-C., and Thiery, J. P. (2013). Translating metastasis-related biomarkers to the clinic – progress and pitfalls. *Nature Reviews Clinical Oncology*, 10(3):169–179. (Cited on page 10.)

[Bidard et al., 2016] Bidard, F.-C., Proudhon, C., and Pierga, J.-Y. (2016). Circulating tumor cells in breast cancer. *Molecular Oncology*. (Cited on page 10.)

[Bidard et al., 2008] Bidard, F.-C., Vincent-Salomon, A., Sigal-Zafrani, B., Dieras, V., Mathiot, C., Mignot, L., Thiery, J.-P., Sastre-Garau, X., and Pierga, J.-Y. (2008). Prognosis of women with stage iv breast cancer depends on detection of circulating tumor cells rather than disseminated tumor cells. *Annals of Oncology*, page mdm507. (Cited on page 9.)

[Biernacki et al., 2003] Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575. (Cited on page 104.)

[Bishop, 2006a] Bishop, C. (2006a). *Pattern recognition and machine learning*. Springer, New York. (Cited on page 77.)

[Bishop, 2006b] Bishop, C. (2006b). *Pattern Recognition and Machine Learning*. Springer. (Cited on pages 96 and 98.)

[Bolón-Canedo et al., 2010] Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2010). On the effectiveness of discretization on gene selection of microarray data. In *Neural networks (ijcnn), the 2010 international joint conference on*, pages 1–8. IEEE. (Cited on page 16.)

[Bolshakova et al., 2006] Bolshakova, N., Azuaje, F., and Cunningham, P. (2006). Incorporating biological domain knowledge into cluster validity assessment. In *Applications of Evolutionary Computing*, pages 13–22. Springer. (Cited on page 32.)

[Borg and Groenen, 2005] Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media. (Cited on page 27.)

[Botteri et al., 2010] Botteri, E., Sandri, M. T., Bagnardi, V., Munzone, E., Zorzino, L., Rotmensz, N., Casadio, C., Cassatella, M. C., Esposito, A., Curigliano, G., et al. (2010). Modeling the relationship between circulating tumour cells number and prognosis of metastatic breast cancer. *Breast cancer research and treatment*, 122(1):211–217. (Cited on page 9.)

[Box and Cox, 1964] Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252. (Cited on page 16.)

[Breiman, 1996] Breiman, L. (1996). Bagging predictors. *Machine learning*. (Cited on page 77.)

[Breiman, 2001] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32. (Cited on pages 26 and 82.)

[Brown and Botstein, 1999] Brown, P. O. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature genetics*, 21:33–37. (Cited on page 13.)

[Brown, 2006] Brown, T. (2006). *Genomes 3*. Garland Science. (Cited on pages 3 and 6.)

[Burnham and Anderson, 2004] Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304. (Cited on page 105.)

[Caldon et al., 2006] Caldon, C. E., Daly, R. J., Sutherland, R. L., and Musgrove, E. A. (2006). Cell cycle control in breast cancer cells. *Journal of cellular biochemistry*, 97(2):261–274. (Cited on page 55.)

[Can et al., 2005] Can, T., Çamoglu, O., and Singh, A. K. (2005). Analysis of protein-protein interaction networks using random walks. In *Proceedings of the 5th international workshop on Bioinformatics*, pages 61–68. ACM. (Cited on page 76.)

[Cantley et al., 1991] Cantley, L. C., Auger, K. R., Carpenter, C., Duckworth, B., Graziani, A., Kapeller, R., and Soltoff, S. (1991). Oncogenes and signal transduction. *Cell*, 64(2):281–302. (Cited on page 45.)

[Carmona-Saez et al., 2006] Carmona-Saez, P., Pascual-Marqui, R. D., Tirado, F., Carazo, J. M., and Pascual-Montano, A. (2006). Biclustering of gene expression data by non-smooth non-negative matrix factorization. *BMC bioinformatics*, 7(1):1. (Cited on page 21.)

[Casella and Berger, 2001] Casella, G. and Berger, R. L. (2001). *Statistical inference*. Duxbury Pacific Grove, CA. (Cited on page 17.)

[Cattell, 1966] Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276. (Cited on page 27.)

[Chaffer and Weinberg, 2011] Chaffer, C. L. and Weinberg, R. A. (2011). A perspective on cancer cell metastasis. *Science*, 331(6024):1559–1564. (Cited on pages 9 and 11.)

[Chambers et al., 2002] Chambers, A. F., Groom, A. C., and MacDonald, I. C. (2002). Metastasis: dissemination and growth of cancer cells in metastatic sites. *Nature Reviews Cancer*, 2(8):563–572. (Cited on page 36.)

[Chen et al., 2011] Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., and Liu, C. (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one*, 6(2):e17238. (Cited on pages 29 and 30.)

[Chen and Wang, 2009] Chen, X. and Wang, L. (2009). Integrating Biological Knowledge with Gene Expression Profiles for Survival Prediction of Cancer. *Journal of Computational Biology*, 16(2):265–278. (Cited on page 33.)

[Chlis et al., 2015] Chlis, N.-K., Sfakianakis, S., Bei, E. S., Iliopoulou, D., Kafetzopoulos, D., and Zervakis, M. (2015). 6th european conference of the international federation for medical and biological engineering: Mbec 2014, 7-11 september 2014, dubrovnik, croatia. pages 594–597, Cham. Springer International Publishing. (Cited on page 129.)

[Chlis et al., 2013] Chlis, N. K., Sfakianakis, S., Bei, E. S., and Zervakis, M. (2013). A generic framework for the elicitation of stable and reliable gene expression signatures. In *Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on*, pages 1–6. (Cited on page 129.)

[Choi et al., 2012] Choi, J. H., Shin, N. R., Moon, H. J., Kwon, C. H., Kim, G. H., Song, G. A., Jeon, T. Y., Kim, D. H., Kim, D. H., and Park, D. Y. (2012). Identification of S100A8 and S100A9 as negative regulators for lymph node metastasis of gastric adenocarcinoma. *Histology and histopathology*, 27(11):1439–1448. (Cited on pages 50 and 57.)

[Choi et al., 2003] Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics (Oxford, England)*, 19 Suppl 1:i84–90. (Cited on page 47.)

[Chuang et al., 2007] Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3. (Cited on pages 33, 37, 66, 71 and 87.)

[Chung, 2007] Chung, F. (2007). The heat kernel as the pagerank of a graph. *Proceedings of the National Academy of Sciences of the United States of America*, 104(50):19735–19740. (Cited on page 76.)

[Churchill, 2002] Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*, 32(Supp):490–495. (Cited on page 29.)

[Clarke et al., 2008] Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., and Wang, Y. (2008). The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8(1):37–49. (Cited on page 3.)

[Crick et al., 1970] Crick, F. et al. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563. (Cited on page 5.)

[Cristofanilli et al., 2007] Cristofanilli, M., Broglio, K. R., Guarneri, V., Jackson, S., Fritsche, H. A., Islam, R., Dawood, S., Reuben, J. M., Kau, S.-W., Lara, J. M., et al. (2007). Circulating tumor cells in metastatic breast cancer: biologic staging beyond tumor burden. *Clinical breast cancer*, 7(6):34–42. (Cited on page 11.)

[Cristofanilli et al., 2004] Cristofanilli, M., Budd, G. T., Ellis, M. J., Stopeck, A., Matera, J., Miller, M. C., Reuben, J. M., Doyle, G. V., Allard, W. J., Terstappen, L. W., et al. (2004). Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *New England Journal of Medicine*, 351(8):781–791. (Cited on pages 9 and 11.)

[Dahm, 2008] Dahm, R. (2008). Discovering DNA: Friedrich miescher and the early years of nucleic acid research. *Human genetics*, 122(6):565–581. (Cited on page 4.)

[Das and Yu, 2012] Das, J. and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Systems Biology*, 6(1):92. (Cited on pages 65 and 75.)

[Dasgupta and Gupta, 2003] Dasgupta, S. and Gupta, A. (2003). An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65. (Cited on page 28.)

[Datta and Datta, 2006] Datta, S. and Datta, S. (2006). Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC bioinformatics*, 7(1):397. (Cited on page 100.)

[Desmedt et al., 2008] Desmedt, C., Haibe-Kains, B., Wirapati, P., Buyse, M., Larsimont, D., Bontempi, G., Delorenzi, M., Piccart, M., and Sotiriou, C. (2008). Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical Cancer Research*, 14(16):5158–5165. (Cited on page 52.)

[Dietterich, 2000] Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, pages 1–15. (Cited on page 77.)

[Ding et al., 2010] Ding, L., Ellis, M. J., Li, S., Larson, D. E., Chen, K., Wallis, J. W., Harris, C. C., McLellan, M. D., Fulton, R. S., Fulton, L. L., et al. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*, 464(7291):999–1005. (Cited on page 43.)

[Dirix et al., 2005] Dirix, L., Van Dam, P., and Vermeulen, P. (2005). Genomics and circulating tumor cells: promising tools for choosing and monitoring adjuvant therapy in patients with early breast cancer? *Current opinion in oncology*, 17(6):551–558. (Cited on page 42.)

[Domingos, 2012] Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87. (Cited on page 22.)

[Dong et al., 2013] Dong, X., Alpaugh, R. K., and Cristofanilli, M. (2013). Circulating tumor cells (CTCs) in breast cancer: a diagnostic tool for prognosis and molecular analysis. *Chinese Journal of Cancer Research*, 24(4):388–398. (Cited on page 64.)

[Drăghici, 2011] Drăghici, S. (2011). *Statistics and Data Analysis for Microarrays Using R and Bioconductor*. Chapman and Hall/CRC, 2nd edition. (Cited on page 19.)

[Dudoit et al., 2002a] Dudoit, S., Fridlyand, J., and Speed, T. (2002a). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87. (Cited on page 21.)

[Dudoit et al., 2002b] Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002b). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica*, pages 111–139. (Cited on page 17.)

[Eckart and Young, 1936] Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218. (Cited on page 118.)

[Edgar et al., 2002a] Edgar, R., Domrachev, M., and Lash, A. E. (2002a). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210. (Cited on pages 15 and 30.)

[Edgar et al., 2002b] Edgar, R., Domrachev, M., and Lash, A. E. (2002b). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210. (Cited on page 46.)

[Ein-Dor et al., 2005] Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178. (Cited on page 37.)

[Ein-Dor et al., 2006] Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, 103(15):5923–5928. (Cited on page 25.)

[Eisen et al., 1998] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868. (Cited on page 20.)

[Enerly et al., 2011] Enerly, E., Steinfeld, I., Kleivi, K., Leivonen, S.-K., Aure, M. R., Russnes, H. G., Rønneberg, J. A., Johnsen, H., Navon, R., Rødland, E., et al. (2011). mirna-mrna integrated analysis reveals roles for mirnas in primary breast tumors. *PloS one*, 6(2):e16915. (Cited on page 47.)

[Ezkurdia et al., 2014] Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human molecular genetics*, 23(22):5866–5878. (Cited on page 6.)

[Fan et al., 2006] Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., van't Veer, L. J., and Perou, C. M. (2006). Concordance among gene-expression–based predictors for breast cancer. *New England Journal of Medicine*, 355(6):560–569. (Cited on page 55.)

[Fawcett, 2006] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874. (Cited on page 81.)

[Fayyad and Irani, 1993] Fayyad, U. and Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *International Joint Conference on Uncertainty in AI*, pages 1022–1027. (Cited on page 16.)

[Fidler, 2002] Fidler, I. J. (2002). The organ microenvironment and cancer metastasis. *Differentiation*, 70(9-10):498–505. (Cited on page 9.)

[Fraley and Raftery, 1999] Fraley, C. and Raftery, A. (1999). MCLUST: Software for model-based cluster analysis. *Journal of Classification*, 16(2):297–306. (Cited on page 99.)

[Franken et al., 2012] Franken, B., de Groot, M. R., Mastboom, W. J., Vermes, I., van der Palen, J., Tibbe, A. G., and Terstappen, L. W. (2012). Circulating tumor cells, disease recurrence and survival in newly diagnosed breast cancer. *Breast Cancer Research*, 14(5):1. (Cited on page 12.)

[Freund and Schapire, 1997] Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139. (Cited on page 77.)

[Friedman et al., 2010] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1):1–968. (Cited on page 68.)

[Ge et al., 2003] Ge, H., Walhout, A. J., and Vidal, M. (2003). Integrating "omic" information: a bridge between genomics and systems biology. *TRENDS in Genetics*, 19(10):551–560. (Cited on page 2.)

[Gentleman et al., 2004] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80. (Cited on page 48.)

[Giuliano et al., 2014] Giuliano, M., Giordano, A., Jackson, S., De Giorgi, U., Mego, M., Cohen, E. N., Gao, H., Anfossi, S., Handy, B. C., Ueno, N. T., et al. (2014). Circulating tumor cells as early predictors of metastatic spread in breast cancer patients with limited metastatic dissemination. *Breast Cancer Research*, 16(5):1. (Cited on page 12.)

[Goh et al., 2007] Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabasi, A. L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690. (Cited on pages 33, 34 and 73.)

[Golub and Van Loan, 2012] Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press. (Cited on page 119.)

[Golub et al., 1999] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537. (Cited on pages 17 and 21.)

[Gradilone et al., 2011] Gradilone, A., Naso, G., Raimondi, C., Cortesi, E., Gandini, O., Vincenzi, B., Saltarelli, R., Chiapparino, E., Spremberg, F., Cristofanilli, M., et al. (2011). Circulating tumor cells (ctcs) in metastatic breast cancer (mbc): prognosis, drug resistance and phenotypic characterization. *Annals of Oncology*, 22(1):86–92. (Cited on page 12.)

[Green et al., 2012] Green, S., Benedetti, J., Smith, A., and Crowley, J. (2012). *Clinical trials in oncology*, volume 28. CRC press. (Cited on page 9.)

[Greene et al., 2012] Greene, B. T., Hughes, A. D., and King, M. R. (2012). Circulating Tumor Cells: The Substrate of Personalized Medicine? *Frontiers in oncology*, 2. (Cited on page 64.)

[Guyon and Elisseeff, 2003] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182. (Cited on page 25.)

[Guyon et al., 2002] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422. (Cited on page 26.)

[Gypas et al., 2011] Gypas, F., Bei, E. S., Zervakis, M., and Sfakianakis, S. (2011). A disease annotation study of gene signatures in a breast cancer microarray dataset. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 5551–5554. (Cited on page 129.)

[Haakensen et al., 2010] Haakensen, V. D., Biong, M., Lingjærde, O. C., Holmen, M. M., Frantzen, J. O., Chen, Y., Navjord, D., Romundstad, L., Lüders, T., Bukholm, I. K., et al. (2010). Expression levels of uridine 5'-diphospho-glucuronosyltransferase genes in breast tissue from healthy women are associated with mammographic density. *Breast Cancer Research*, 12(4):R65. (Cited on page 47.)

[Hall, 1999] Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato. (Cited on page 25.)

[Hanahan and Weinberg, 2000] Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *cell*, 100(1):57–70. (Cited on pages 53 and 55.)

[Hanahan and Weinberg, 2011] Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, 144(5):646–674. (Cited on pages 53 and 55.)

[Hanisch et al., 2002] Hanisch, D., Zien, A., Zimmer, R., and Lengauer, T. (2002). Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18(suppl_1):145–154. (Cited on page 34.)

[Hansen, 1998] Hansen, P. (1998). *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. Society for Industrial Mathematics. (Cited on page 108.)

[Harris et al., 2012] Harris, J. R., Lippman, M. E., Osborne, C. K., and Morrow, M. (2012). *Diseases of the Breast*. Lippincott Williams & Wilkins. (Cited on page 8.)

[Hartigan, 1975] Hartigan, J. A. (1975). Clustering algorithms. (Cited on page 20.)

[Hayes, 2011] Hayes, B. (2011). An adventure in the Nth dimension. *American Scientist*, 99(6):442–446. (Cited on pages 22 and 23.)

[Hayes et al., 2006] Hayes, D. F., Cristofanilli, M., Budd, G. T., Ellis, M. J., Stopeck, A., Miller, M. C., Matera, J., Allard, W. J., Doyle, G. V., and Terstappen, L. W. (2006). Circulating tumor cells at each follow-up time point during therapy of metastatic breast cancer patients predict progression-free and overall survival. *Clinical Cancer Research*, 12(14):4218–4224. (Cited on page 9.)

[He and Niyogi, 2004] He, X. and Niyogi, P. (2004). Locality preserving projections. In Thrun, S., Saul, L. K., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*, pages 153–160. MIT Press. (Cited on page 27.)

[Hernandez et al., 2006] Hernandez, L. M., Blazer, D. G., et al. (2006). *Genes, Behavior, and the Social Environment:: Moving Beyond the Nature/Nurture Debate*. National Academies Press. (Cited on page 6.)

[Hoeflich et al., 2009] Hoeflich, K. P., O'Brien, C., Boyd, Z., Cavet, G., Guerrero, S., Jung, K., Januario, T., Savage, H., Punnoose, E., Truong, T., et al. (2009). In vivo antitumor activity of mek and phosphatidylinositol 3-kinase inhibitors in basal-like breast cancer models. *Clinical Cancer Research*, 15(14):4649–4664. (Cited on page 47.)

[Hofree et al., 2013] Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nature Methods*, 10(11):1108–1115. (Cited on page 92.)

[Hoh and Ott, 2004] Hoh, J. and Ott, J. (2004). Genetic dissection of diseases: design and methods. *Current opinion in genetics & development*, 14(3):229–232. (Cited on page 2.)

[Hong et al., 2006] Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., and Chory, J. (2006). Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827. (Cited on page 19.)

[Huang and Pan, 2006] Huang, D. and Pan, W. (2006). Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics*, 22(10):1259–1268. (Cited on page 35.)

[Huang et al., 2003] Huang, E., Cheng, S., Dressman, H., Pittman, J., Tsou, M., Horng, C., Bild, A., Iversen, E., Liao, M., Chen, C., et al. (2003). Gene expression predictors of breast cancer outcomes. *The Lancet*, 361(9369):1590–1596. (Cited on pages 98, 102 and 105.)

[Hung, 2013] Hung, J.-H. (2013). Gene set/pathway enrichment analysis. *Data Mining for Systems Biology: Methods and Protocols*, pages 201–213. (Cited on page 55.)

[Hunter and Alsarraj, 2009] Hunter, K. W. and Alsarraj, J. (2009). Gene expression profiles and breast cancer metastasis: a genetic perspective. *Clinical & experimental metastasis*, 26(6):497–503. (Cited on pages 45 and 57.)

[Ideker et al., 2001] Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, 2(1):343–372. (Cited on page 107.)

[Ideker et al., 2002] Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl 1):S233–S240. (Cited on page 33.)

[Ideker and Sharan, 2008] Ideker, T. and Sharan, R. (2008). Protein networks in disease. *Genome research*, 18(4):644–652. (Cited on pages 63 and 64.)

[Ignatiadis et al., 2015] Ignatiadis, M., Lee, M., and Jeffrey, S. S. (2015). Circulating tumor cells and circulating tumor DNA: Challenges and opportunities on the path to clinical utility. *Clinical Cancer Research*, 21(21):4786–4800. (Cited on page 35.)

[Irizarry et al., 2003] Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of affymetrix genechip probe level data. *Nucleic acids research*, 31(4):e15–e15. (Cited on page 14.)

[Jaffrézic et al., 2007] Jaffrézic, F., Marot, G., Degrelle, S., Hue, I., and Foulley, J.-L. (2007). A structural mixed model for variances in differential gene expression studies. *Genetical research*, 89(1):19–25. (Cited on page 47.)

[Janni et al., 2016] Janni, W. J., Rack, B., Terstappen, L. W., Pierga, J.-Y., Taran, F.-A., Fehm, T., Hall, C., de Groot, M. R., Bidard, F.-C., Friedl, T. W., et al. (2016). Pooled analysis of the prognostic relevance of circulating tumor cells in primary breast cancer. *Clinical Cancer Research*, 22(10):2583–2593. (Cited on page 12.)

[Jelizarow et al., 2010] Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K., and Boulesteix, A. (2010). Over-optimism in bioinformatics: an illustration. *Bioinformatics (Oxford, England)*, pages btq323+. (Cited on page 107.)

[Johnson et al., 2007] Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostat*, 8(1):118–127. (Cited on pages 30 and 48.)

[Jolie, 2013] Jolie, A. (2013). My medical choice. *The New York Times*, 14(05):2013. (Cited on page 9.)

[Jonsson and Bates, 2006] Jonsson, P. F. and Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18):2291–2297. (Cited on pages 64 and 73.)

[Joosse et al., 2015] Joosse, S. A., Gorges, T. M., and Pantel, K. (2015). Biology, detection, and clinical implications of circulating tumor cells. *EMBO molecular medicine*, 7(1):1–11. (Cited on page 12.)

[Kallergi et al., 2015] Kallergi, G., Tsintari, V., Sfakianakis, S., Zervakis, M., Mavroudis, D., and Georgoulias, V. (2015). CXCR4 pathways in CTCs: from bioinformatics to immunophenotype. *Cancer Research*, 75(15 Supplement):1592. (Cited on pages 114 and 127.)

[Kandula et al., 2012] Kandula, M., Ch, K. K., Kanth, R., VV, L. A., Murthy, S., and Raju, Y. A. (2012). Differences in gene expression profiles between human breast tissue and peripheral blood samples for breast cancer detection. *Journal of Cancer Science & Therapy*, 2012. (Cited on page 45.)

[Kanehisa and Goto, 2000] Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27. (Cited on page 94.)

[Kaufman and Rousseeuw, 2009] Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons. (Cited on page 20.)

[Kerr and Churchill, 2001] Kerr, M. K. and Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics*, 2(2):183–201. (Cited on page 19.)

[Khatri and Drăghici, 2005] Khatri, P. and Drăghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595. (Cited on page 32.)

[Kim and Kim, 2008] Kim, S.-Y. and Kim, Y. S. (2008). A gene sets approach for identifying prognostic gene signatures for outcome prediction. *BMC Genomics*, 9(1):177. (Cited on page 56.)

[Kittler et al., 1998] Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On Combining Classifiers. *IEEE Trans. Pattern Anal. Mach. Intell. ()*, 20(3):226–239. (Cited on page 77.)

[Kohavi and John, 1997] Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*. (Cited on page 74.)

[Kou et al., 1981] Kou, L., Markowsky, G., and Berman, L. (1981). A fast algorithm for Steiner trees. *Acta informatica*. (Cited on page 65.)

[Kustra and Zagdanski, 2006] Kustra, R. and Zagdanski, A. (2006). Incorporating gene ontology in clustering gene expression data. In *19th IEEE International Symposium on Computer-Based Medical Systems, 2006. CBMS 2006*, pages 555–563. (Cited on page 34.)

[Labelle and Hynes, 2012] Labelle, M. and Hynes, R. O. (2012). The initial hours of metastasis: the importance of cooperative host–tumor cell interactions during hematogenous dissemination. *Cancer discovery*, 2(12):1091–1099. (Cited on pages 35 and 36.)

[LaBreche et al., 2011] LaBreche, H. G., Nevins, J. R., and Huang, E. (2011). Integrating factor analysis and a transgenic mouse model to reveal a peripheral blood predictor of breast tumors. *BMC medical genomics*, 4(1):61. (Cited on page 47.)

[Lakhtakia, 2014] Lakhtakia, R. (2014). A brief history of breast cancer: Part i: Surgical domination reinvented. *Sultan Qaboos University medical journal*, 14(2):e166. (Cited on page 7.)

[Lang et al., 2015] Lang, J. E., Scott, J. H., Wolf, D. M., Novak, P., Punj, V., Magbanua, M. J. M., Zhu, W., Mineyev, N., Haqq, C. M., Crothers, J. R., Esserman, L. J., Tripathy, D., van t Veer, L., and Park, J. W. (2015). Expression profiling of circulating tumor cells in metastatic breast cancer. *Breast cancer research and treatment*, 149(1):121–131. (Cited on pages 80, 86, 87, 88, 91 and 109.)

[Ledoit and Wolf, 2004] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411. (Cited on page 26.)

[Lee et al., 2008] Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., and Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS comput biol*, 4(11):e1000217. (Cited on page 37.)

[Lee and Verleysen, 2007] Lee, J. A. and Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Springer Science & Business Media. (Cited on page 27.)

[Leek et al., 2010] Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739. (Cited on pages 29 and 59.)

[Leiserson et al., 2014] Leiserson, M. D. M., Vandin, F., Wu, H., Dobson, J. R., Eldridge, J. V., Thomas, J. L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., Lawrence, M. S., Gonzalez-Perez, A., Tamborero, D., Cheng, Y., Ryslik, G. A., Lopez-Bigas, N., Getz, G., and Ding, L Raphael, B. J. (2014). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, pages 1–11. (Cited on page 92.)

[Li et al., 2010] Li, Y., Liu, L., Bai, X., Cai, H., Ji, W., Guo, D., and Zhu, Y. (2010). Comparative study of discretization methods of microarray data for inferring transcriptional regulatory networks. *BMC bioinformatics*, 11(1):520. (Cited on page 16.)

[Lianidou, 2014] Lianidou, E. S. (2014). Circulating Tumor Cells: A Non-invasive Liquid Biopsy in Cancer. In *Molecular Testing in Cancer*, pages 119–132. Springer New York, New York, NY. (Cited on page 35.)

[Liew et al., 2011] Liew, A. W.-C., Law, N.-F., and Yan, H. (2011). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in bioinformatics*, 12(5):498–513. (Cited on page 15.)

[Liu et al., 2011] Liu, R.-Z., Graham, K., Glubrecht, D. D., Germain, D. R., Mackey, J. R., and Godbout, R. (2011). Association of fabp5 expression with poor survival in triple-negative breast cancer: implication for retinoic acid therapy. *The American journal of pathology*, 178(3):997–1008. (Cited on page 47.)

[Lockhart and Winzeler, 2000] Lockhart, D. J. and Winzeler, E. A. (2000). Genomics, gene expression and dna arrays. *Nature*, 405(6788):827–836. (Cited on page 13.)

[Lopez et al., 2012] Lopez, F. J., Cuadros, M., Cano, C., Concha, A., and Blanco, A. (2012). Biomedical application of fuzzy association rules for identifying breast cancer biomarkers. *Medical & biological engineering & computing*, 50(9):981–990. (Cited on page 56.)

[Lovasz, 1993] Lovasz, L. (1993). Random walks on graphs: A survey. *Combinatorics*. (Cited on page 76.)

[Lozy and Karantza, 2012] Lozy, F. and Karantza, V. (2012). Autophagy and cancer cell metabolism. In *Seminars in cell & developmental biology*, volume 23, pages 395–401. Elsevier. (Cited on page 53.)

[Lucci et al., 2012] Lucci, A., Hall, C. S., Lodhi, A. K., Bhattacharyya, A., Anderson, A. E., Xiao, L., Bedrosian, I., Kuerer, H. M., and Krishnamurthy, S. (2012). Circulating tumour cells in non-metastatic breast cancer: a prospective study. *The lancet oncology*, 13(7):688–695. (Cited on page 12.)

[Lum et al., 2013] Lum, P., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., and Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. *Scientific reports*, 3. (Cited on page 27.)

[Madeira and Oliveira, 2004] Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 1(1):24–45. (Cited on page 21.)

[Magbanua et al., 2015] Magbanua, M. J. M., Carey, L. A., DeLuca, A., Hwang, J., Scott, J. H., Rimawi, M. F., Mayer, E. L., Marcom, P. K., Liu, M. C., Esteva, F. J., et al. (2015). Circulating tumor cell analysis in metastatic triple-negative breast cancers. *Clinical Cancer Research*, 21(5):1098–1105. (Cited on page 35.)

[Maltoni et al., 2015] Maltoni, R., Fici, P., Amadori, D., Gallerani, G., Cocchi, C., Zoli, M., Rocca, A., Cecconetto, L., Folli, S., Scarpi, E., et al. (2015). Circulating tumor cells in early breast cancer: a connection with vascular invasion. *Cancer letters*, 367(1):43–48. (Cited on page 12.)

[Markiewicz et al., 2014] Markiewicz, A., Książkiewicz, M., Wełnicka-Jaśkiewicz, M., Seroczyńska, B., Skokowski, J., Szade, J., and Żaczek, A. J. (2014). Mesenchymal phenotype of CTC-enriched blood fraction and lymph node metastasis formation potential. *PloS one*, 9(4):e93901. (Cited on page 36.)

[Marot et al., 2009] Marot, G., Foulley, J.-L., Mayer, C.-D., and Jaffrézic, F. (2009). Moderated effect size and P-value combinations for microarray meta-analyses. *Bioinformatics*, 25(20):2692–2699. (Cited on page 47.)

[Mason and Verwoerd, 2007] Mason, O. and Verwoerd, M. (2007). Graph theory and networks in biology. *Systems Biology, IET*, 1(2):89–119. (Cited on page 32.)

[Massagué and Obenauf, 2016] Massagué, J. and Obenauf, A. C. (2016). Metastatic colonization by circulating tumour cells. *Nature*, 529(7586):298–306. (Cited on page 9.)

[McInnes et al., 2015] McInnes, L. M., Jacobson, N., Redfern, A., Dowling, A., Thompson, E. W., and Saunders, C. M. (2015). Clinical implications of circulating tumor cells of breast cancer patients: role of epithelial–mesenchymal plasticity. *Cellular and Phenotypic Plasticity in Cancer*, page 18. (Cited on page 12.)

[McLachlan and Krishnan, 1997] McLachlan, G. and Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley New York. (Cited on page 96.)

[McLachlan and Peel, 2000] McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley-Interscience. (Cited on pages 95, 98 and 104.)

[Mego et al., 2016] Mego, M., Cholujova, D., Minarik, G., Sedlackova, T., Gronesova, P., Karaba, M., Benca, J., Cingelova, S., Cierna, Z., Manasova, D., et al. (2016). Cxcr4-sdf-1 interaction potentially mediates trafficking of circulating tumor cells in primary breast cancer. *BMC cancer*, 16(1):1. (Cited on page 50.)

[Meyer, 2000] Meyer, C. D. (2000). *Matrix analysis and applied linear algebra*, volume 2. Siam. (Cited on pages 27 and 117.)

[Mi and Thomas, 2009] Mi, H. and Thomas, P. (2009). PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Protein Networks and Pathway Analysis*, pages 123–140. (Cited on page 86.)

[Miller et al., 2009] Miller, M. C., Doyle, G. V., and Terstappen, L. W. (2009). Significance of circulating tumor cells detected by the cellsearch system in patients with metastatic breast colorectal and prostate cancer. *Journal of oncology*, 2010. (Cited on page 35.)

[Molloy et al., 2012] Molloy, T. J., Roepman, P., Naume, B., and Veer, L. J. v. (2012). A Prognostic Gene Expression Profile That Predicts Circulating Tumor Cell Presence in Breast Cancer Patients. *PloS one*, 7(2):e32426. (Cited on pages 37, 46, 47, 49, 50, 55 and 58.)

[Monti et al., 2003] Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118. (Cited on page 21.)

[Mosca et al., 2010a] Mosca, E., Alfieri, R., Merelli, I., Viti, F., Calabria, A., and Milanesi, L. (2010a). A multilevel data integration resource for breast cancer study. *BMC Systems Biology*, 4(1):76. (Cited on page 84.)

[Mosca et al., 2010b] Mosca, E., Alfieri, R., Merelli, I., Viti, F., Calabria, A., and Milanesi, L. (2010b). A multilevel data integration resource for breast cancer study. *BMC Systems Biology*, 4(1):76. (Cited on page 52.)

[Murphy, 2012] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective.* MIT press. (Cited on page 21.)

[Nadal et al., 2013] Nadal, R., Lorente, J. A., Rosell, R., and Serrano, M. J. (2013). Relevance of molecular characterization of circulating tumor cells in breast cancer in the era of targeted therapies. *Expert review of molecular diagnostics*, 13(3):295–307. (Cited on page 36.)

[Nagrath et al., 2007] Nagrath, S., Sequist, L. V., Maheswaran, S., Bell, D. W., Irimia, D., Ulkus, L., Smith, M. R., Kwak, E. L., Digumarthy, S., Muzikansky, A., et al. (2007). Isolation of rare circulating tumour cells in cancer patients by microchip technology. *Nature*, 450(7173):1235–1239. (Cited on page 35.)

[Notas et al., 2015] Notas, G., Pelekanou, V., Kampa, M., Alexakis, K., Sfakianakis, S., Laliotis, A., Askoxilakis, J., Tsentelierou, E., Tzardi, M., Tsapis, A., and Castanas, E. (2015). Tamoxifen induces a pluripotency signature in breast cancer cells and human tumors. *Molecular Oncology*, 9(9):1744 – 1759. (Cited on pages 18 and 128.)

[Obermayr et al., 2010] Obermayr, E., Cabo, F. S., Tea, M. K., Singer, C., Krainer, M., Fischer, M., Sehouli, J., Reinthaller, A., Horvat, R., Heinze, G., Tong, D., and Zeillinger, R. (2010). Assessment of a six gene panel for the molecular detection of circulating tumor cells in the blood of female cancer patients. *BMC Cancer*, 10(1):666–666. (Cited on pages 36, 46 and 59.)

[Ojala and Garriga, 2009] Ojala, M. and Garriga, G. (2009). Permutation Tests for Studying Classifier Performance. In *Ninth International Conference on Data Mining*, pages 908–913. IEEE Computer Society. (Cited on page 68.)

[Oliveros, 2007] Oliveros, J. C. (2007). Venny, an interactive tool for comparing lists with venn diagrams, `http://bioinfogp.cnb.csic.es/tools/venny/index.html`. (Cited on page 51.)

[Osborne et al., 2009] Osborne, J. D., Flatow, J., Holko, M., Lin, S. M., Kibbe, W. A., Zhu, L. J., Danila, M. I., Feng, G., and Chisholm, R. L. (2009). Annotating the human genome with disease ontology. *BMC genomics*, 10(Suppl 1):S6. (Cited on page 86.)

[Ouellet et al., 2011] Ouellet, V., Tiedemann, K., Mourskaia, A., Fong, J. E., Tran-Thanh, D., Amir, E., Clemons, M., Perbal, B., Komarova, S. V., and Siegel, P. M. (2011). Ccn3 impairs osteoblast and stimulates osteoclast differentiation to favor breast cancer metastasis to bone. *The American journal of pathology*, 178(5):2377–2388. (Cited on page 55.)

[Pachmann et al., 2014] Pachmann, K., Stein, E., Spitz, G., Schill, E., and Pachmann, U. (2014). Chemosensitivity Testing of Circulating Epithelial Cells (CETC) in Breast Cancer Patients and Correlation to Clinical Outcome. *Cancer Research*, 69(24 Supplement):2044–2044. (Cited on page 35.)

[Paik et al., 2004] Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826. (Cited on page 37.)

[Pan, 2006] Pan, W. (2006). Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*, 22(7):795–801. (Cited on pages 107 and 108.)

[Park et al., 2009] Park, S., Holmes-Tisch, A. J., Shim, Y. M., Kim, J., Kim, H. S., Lee, J., Park, Y. H., Ahn, J. S., Park, K., Jänne, P. A., et al. (2009). Discordance of molecular biomarkers associated with epidermal growth factor receptor pathway between primary tumors and lymph node metastasis in non-small cell lung cancer. *Journal of Thoracic Oncology*, 4(7):809–815. (Cited on page 36.)

[Parkinson et al., 2009] Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I., Farne, A., et al. (2009). Arrayexpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic acids research*, 37(suppl 1):D868–D872. (Cited on pages 15 and 30.)

[Paterlini-Brechot and Benali, 2007] Paterlini-Brechot, P. and Benali, N. L. (2007). Circulating tumor cells (CTC) detection: clinical impact and future directions. *Cancer letters*, 253(2):180–204. (Cited on page 35.)

[Pau Ni et al., 2010] Pau Ni, I. B., Zakaria, Z., Muhammad, R., Abdullah, N., Ibrahim, N., Aina Emran, N., Hisham Abdullah, N., and Syed Hussain, S. N. A. (2010). Gene expression patterns distinguish breast carcinomas from normal breast tissues: The Malaysian context. *Pathology - Research and Practice*, 206(4):223–228. (Cited on pages 29 and 47.)

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. (Cited on page 92.)

[Pensa et al., 2004] Pensa, R. G., Leschi, C., Besson, J., and Boulicaut, J.-F. (2004). Assessment of discretization techniques for relevant pattern discovery from gene expression data. In *BIOKDD04: Workshop on Data Mining in Bioinformatics August 22nd, 2004 Seattle, WA, USA*, page 24. Citeseer. (Cited on page 16.)

[Pentney and Meila, 2005] Pentney, W. and Meila, M. (2005). Spectral clustering of biological sequence data. In *AAAI*, volume 5, pages 845–850. (Cited on page 21.)

[Piñero et al., 2015] Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., and Furlong, L. I. (2015). Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015:bav028. (Cited on page 85.)

[Pollen et al., 2014] Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., Li, N., Szpankowski, L., Fowler, B., Chen, P., et al. (2014). Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology*, 32(10):1053–1058. (Cited on page 28.)

[Potamias et al., 2004] Potamias, G., Koumakis, L., and Moustakis, V. (2004). Gene selection via discretized gene-expression profiles and greedy feature-elimination. In *Methods and Applications of Artificial Intelligence*, pages 256–266. Springer. (Cited on page 16.)

[Powell et al., 2012] Powell, A. A., Talasaz, A. H., Zhang, H., Coram, M. A., Reddy, A., Deng, G., Telli, M. L., Advani, R. H., Carlson, R. W., Mollick, J. A., Sheth, S., Kurian, A. W., Ford, J. M., Stockdale, F. E., Quake, S. R., Pease, R. F., Mindrinos, M. N., Bhanot, G., Dairkee, S. H., Davis, R. W., and Jeffrey, S. S. (2012). Single Cell Profiling of Circulating Tumor Cells: Transcriptional Heterogeneity and Diversity from Breast Cancer Cell Lines. *PloS one*, 7(5):e33788. (Cited on pages 49 and 50.)

[Powers, 2011] Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63. (Cited on page 78.)

[Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106. (Cited on page 25.)

[Radich et al., 2004] Radich, J. P., Mao, M., Stepaniants, S., Biery, M., Castle, J., Ward, T., Schimmack, G., Kobayashi, S., Carleton, M., Lampe, J., et al. (2004). Individual-specific variation of gene expression in peripheral blood leukocytes. *Genomics*, 83(6):980–988. (Cited on page 45.)

[Rapaport et al., 2007] Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., and Vert, J.-P. (2007). Classification of microarray data using gene networks. *BMC bioinformatics*, 8(1):35. (Cited on page 33.)

[Rhodes and Chinnaiyan, 2005] Rhodes, D. R. and Chinnaiyan, A. M. (2005). Integrative analysis of the cancer transcriptome. *Nature genetics*, 37:S31–S37. (Cited on page 30.)

[Riethdorf and Pantel, 2010] Riethdorf, S. and Pantel, K. (2010). Advancing personalized cancer therapy by detection and characterization of circulating carcinoma cells. *Annals of the New York Academy of Sciences*, 1210(1):66–77. (Cited on page 42.)

[Rogers et al., 2005] Rogers, S., Williams, R. D., and Campbell, C. (2005). Class prediction with microarray datasets. In *Bioinformatics Using Computational Intelligence Paradigms*, pages 119–141. Springer. (Cited on page 21.)

[Rosenlicht, 1986] Rosenlicht, M. (1986). *Introduction to analysis*. Dover, New York. (Cited on page 78.)

[Roweis and Saul, 2000] Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326. (Cited on page 27.)

[Ruschhaupt et al., 2004] Ruschhaupt, M., Huber, W., Poustka, A., Mansmann, U., et al. (2004). A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1078. (Cited on page 106.)

[Saeys et al., 2007] Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507. (Cited on pages 25 and 94.)

[Salhia et al., 2014] Salhia, B., Kiefer, J., Ross, J. T. D., Metapally, R., Martinez, R. A., Johnson, K. N., DiPerna, D. M., Paquette, K. M., Jung, S., Nasser, S., Wallstrom, G., Tembe, W., Baker, A., Carpten, J., Resau, J., Ryken, T., Sibenaller, Z., Petricoin, E. F., Liotta, L. A., Ramanathan, R. K., Berens, M. E., and Tran, N. L. (2014). Integrated genomic and epigenomic analysis of breast cancer brain metastasis. *PloS one*, 9(1):e85448. (Cited on pages 68, 71 and 102.)

[Schäfer and Strimmer, 2005] Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4. (Cited on page 26.)

[Schapire, 1990] Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, 5(2):197–227. (Cited on page 77.)

[Schobesberger et al., 2008] Schobesberger, M., Baltzer, A., Oberli, A., Kappeler, A., Gugger, M., Burger, H., and Jaggi, R. (2008). Gene expression variation between distinct areas of breast cancer measured from paraffin-embedded tissue cores. *BMC cancer*, 8(1):1. (Cited on page 45.)

[Scott, 2015] Scott, D. W. (2015). *Multivariate Density Estimation*. Theory, Practice, and Visualization. John Wiley & Sons, Inc, Hoboken, NJ. (Cited on page 22.)

[Scott and Thompson, 1983] Scott, D. W. and Thompson, J. R. (1983). Probability density estimation in higher dimensions. In *Computer Science and Statistics: Proceedings of the fifteenth symposium on the interface*, volume 528, pages 173–179. North-Holland, Amsterdam. (Cited on page 23.)

[Scriver and Waters, 1999] Scriver, C. R. and Waters, P. J. (1999). Monogenic traits are not simple: lessons from phenylketonuria. *Trends in genetics*, 15(7):267–272. (Cited on page 2.)

[Segal et al., 2005] Segal, E., Friedman, N., Kaminski, N., Regev, A., and Koller, D. (2005). From signatures to models: understanding cancer using microarrays. *Nature genetics*, 37:S38–S45. (Cited on page 75.)

[Sfakianakis et al., 2016a] Sfakianakis, S., Bei, E. S., and Zervakis, M. (2016a). Exploratory analysis of local gene groups in breast cancer guided by biological networks. *Health and Technology*, xx(xx):xx. (Cited on page 128.)

[Sfakianakis et al., 2016b] Sfakianakis, S., Bei, E. S., and Zervakis, M. (2016b). Stacking of network based classifiers with application in breast cancer classification. In *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*, pages 1079–1084. Springer. (Cited on pages 128 and 129.)

[Sfakianakis et al., 2015] Sfakianakis, S., Bei, E. S., Zervakis, M., and Kafetzopoulos, D. (2015). A network-based approach to enrich gene signatures for the prediction of breast cancer metastases. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 6497–6500. IEEE. (Cited on page 129.)

[Sfakianakis et al., 2014] Sfakianakis, S., Bei, E. S., Zervakis, M., Vassou, D., and Kafetzopoulos, D. (2014). On the Identification of Circulating Tumor Cells in Breast Cancer. *Biomedical and Health Informatics, IEEE Journal of*, 18(3):773–782. (Cited on pages 64, 75, 80, 81, 82, 88, 127, 128 and 129.)

[Sfakianakis et al., 2010a] Sfakianakis, S., Blazantonakis, M., Dimou, I., Zervakis, M., Tsiknakis, M., Potamias, G., Kafetzopoulos, D., and Lowe, D. (2010a). Decision support based on genomics: integration of data- and knowledge-driven reasoning. *International Journal of Biomedical Engineering and Technology*, 3(3-4):287–307. (Cited on pages 2 and 127.)

[Sfakianakis et al., 2010b] Sfakianakis, S., Zervakis, M., Tsiknakis, M., and Kafetzopoulos, D. (2010b). Integration of biological knowledge in the mixture-of-Gaussians analysis of genomic clustering. In *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on*, pages 1–4. IEEE. (Cited on pages 128 and 129.)

[Shabo et al., 2013] Shabo, I., Olsson, H., Stål, O., and Svanvik, J. (2013). Breast cancer expression of dap12 is associated with skeletal and liver metastases and poor survival. *Clinical breast cancer*, 13(5):371–377. (Cited on page 55.)

[Shashirekha and Wani, 2015] Shashirekha, H. and Wani, A. H. (2015). Analysis of imputation algorithms for microarray gene expression data. In *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, pages 589–593. IEEE. (Cited on page 15.)

[Shastry, 1995] Shastry, B. (1995). Overexpression of genes in health and sickness. a bird's eye view. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 112(1):1–13. (Cited on page 45.)

[Shi et al., 2012] Shi, M., Beauchamp, R. D., and Zhang, B. (2012). A Network-Based Gene Expression Signature Informs Prognosis and Treatment for Colorectal Cancer Patients. *PloS one*, 7(7):e41292. (Cited on page 92.)

[Shi et al., 2010] Shi, Z., Derow, C. K., and Zhang, B. (2010). Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. *BMC systems biology*, 4(1):1. (Cited on page 53.)

[Shin et al., 2011] Shin, G., Kang, T.-W., Yang, S., Baek, S.-J., Jeong, Y.-S., and Kim, S.-Y. (2011). Gent: gene expression database of normal and tumor tissues. *Cancer informatics*, 10:149. (Cited on page 45.)

[Siegel et al., 2015] Siegel, R. L., Miller, K. D., and Jemal, A. (2015). Cancer statistics, 2015. *CA: a cancer journal for clinicians*, 65(1):5–29. (Cited on pages 2, 6 and 8.)

[Sieuwerts et al., 2011] Sieuwerts, A. M., Mostert, B., Bolt-de Vries, J., Peeters, D. J., de Jongh, F., Stouthard, J., van Galen, A., Dirix, L. Y., van Dam, P. A., de Weerd, V., et al. (2011). mrna and microrna expression profiles in circulating tumor cells and primary tumors of metastatic breast cancer patients. *Clinical Cancer Research*, pages clincanres–0255. (Cited on page 36.)

[Singh et al., 2002] Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209. (Cited on page 98.)

[Slawski et al., 2008] Slawski, M., Daumer, M., and Boulesteix, A.-L. (2008). CMA–a comprehensive bioconductor package for supervised classification with high dimensional data. *BMC bioinformatics*, 9(1):439. (Cited on page 21.)

[Smyth, 2005] Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer. (Cited on page 18.)

[Sotiriou and Piccart, 2007] Sotiriou, C. and Piccart, M. J. (2007). Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nature Reviews Cancer*, 7(7):545–553. (Cited on page 75.)

[Sotiriou and Pusztai, 2009] Sotiriou, C. and Pusztai, L. (2009). Gene-expression signatures in breast cancer. *N Engl J Med*, 360(8):790–800. (Cited on page 107.)

[Stambuk et al., 2010] Stambuk, S., Sundov, D., Kuret, S., Beljan, R., and Andelinović, S. (2010). Future perspectives of personalized oncology. *Collegium antropologicum*, 34(2):763–769. (Cited on page 71.)

[Stoecklein et al., 2015] Stoecklein, N. H., Fischer, J. C., Niederacher, D., and Terstappen, L. W. (2015). Challenges for ctc-based liquid biopsies: low ctc frequency and diagnostic leukapheresis as a potential solution. *Expert review of molecular diagnostics*, pages 1–18. (Cited on page 35.)

[Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550. (Cited on page 84.)

[Suzuki and Tarin, 2007] Suzuki, M. and Tarin, D. (2007). Gene expression profiling of human lymph node metastases and matched primary breast carcinomas: clinical implications. *Molecular oncology*, 1(2):172–180. (Cited on page 36.)

[Tai and Pan, 2007] Tai, F. and Pan, W. (2007). Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. *Bioinformatics*, 23(14):1775. (Cited on page 107.)

[Taylor et al., 2009] Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J. L. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature biotechnology*, 27(2):199–204. (Cited on page 37.)

[Tenenbaum et al., 2000] Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323. (Cited on page 26.)

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288. (Cited on pages 26 and 94.)

[Tibshirani et al., 2002] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United states of America*, 99(10):6567. (Cited on pages 26 and 94.)

[Tikhonov, 1977] Tikhonov, A. N. (1977). *Solutions of ill-posed problems*. Scripta series in mathematics. Washington : Winston ; New York : distributed solely by Halsted Press. (Cited on page 26.)

[Tripathi et al., 2008] Tripathi, A., King, C., De la Morenas, A., Perry, V. K., Burke, B., Antoine, G. A., Hirsch, E. F., Kavanah, M., Mendez, J., Stone, M., et al. (2008). Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients. *International Journal of Cancer*, 122(7):1557–1566. (Cited on page 47.)

[Tritchler et al., 2009] Tritchler, D., Parkhomenko, E., and Beyene, J. (2009). Filtering genes for cluster and network analysis. *BMC bioinformatics*, 10(1):193. (Cited on page 107.)

[Troyanskaya et al., 2001] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525. (Cited on pages 15 and 48.)

[Tsakaneli et al., 2016] Tsakaneli, S., Bei, E. S., and Zervakis, M. (2016). Comparing genomic network methodologies: A combined approach for cancer prognosis. In *XIV Mediterranean Conference on Medical and Biological*

Engineering and Computing 2016, pages 506–511. Springer. (Cited on page 129.)

[Tsiliki et al., 2011] Tsiliki, G., Zervakis, M., Ioannou, M., Sanidas, E., Stathopoulos, E., Potamias, G., Tsiknakis, M., and Kafetzopoulos, D. (2011). Multi-platform data integration in microarray analysis. *Information Technology in Biomedicine, IEEE Transactions on*, 15(6):806–812. (Cited on pages 30 and 59.)

[Turnbull et al., 2012] Turnbull, A., Kitchen, R., Larionov, A., Renshaw, L., Dixon, J., and Sims, A. (2012). Direct integration of intensity-level data from Affymetrix and Illumina microarrays improves statistical power for robust reanalysis. *BMC Medical Genomics*, 5(1):35. (Cited on page 59.)

[Tusher et al., 2001] Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121. (Cited on pages 18, 19 and 48.)

[Valastyan and Weinberg, 2011] Valastyan, S. and Weinberg, R. A. (2011). Tumor metastasis: molecular insights and evolving paradigms. *Cell*, 147(2):275–292. (Cited on page 2.)

[Van der Laan et al., 2003] Van der Laan, M., Pollard, K., and Bryan, J. (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584. (Cited on page 20.)

[Van der Maaten and Hinton, 2008] Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85. (Cited on page 28.)

[Van't Veer et al., 2002] Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536. (Cited on pages 2 and 37.)

[Vanunu et al., 2010] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating Genes and Protein Complexes with Disease via Network Propagation. *PLoS Comput Biol*, 6(1):e1000641. (Cited on page 74.)

[Vidal et al., 2011] Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). Interactome Networks and Human Disease. *Cell*, 144(6):986–998. (Cited on pages 32 and 74.)

[Wachi et al., 2005] Wachi, S., Yoneda, K., and Wu, R. (2005). Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, 21(23):4205–4208. (Cited on page 64.)

[Wang et al., 2005] Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679. (Cited on page 37.)

[Wang et al., 2014] Wang, L Xiao, Y., Ping, Y., Li, J., Zhao, H., Li, F., Hu, J., Zhang, H., Deng, Y., Tian, J., and Li, X. (2014). Integrating Multi-Omics for Uncovering the Architecture of Cross-Talking Pathways in Breast Cancer. *PloS one*, 9(8):e104282. (Cited on page 92.)

[Watson et al., 1953] Watson, J. D., Crick, F. H., et al. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356):737–738. (Cited on page 4.)

[Wei and Pan, 2008] Wei, P. and Pan, W. (2008). Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics*, 24(3):404. (Cited on page 107.)

[Weigelt et al., 2003] Weigelt, B., Glas, A. M., Wessels, L. F. A., Witteveen, A. T., Peterse, J. L., and van't Veer, L. J. (2003). Gene expression profiles of primary breast tumors maintained in distant metastases. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15901–15905. (Cited on page 57.)

[Weigelt et al., 2005] Weigelt, B., Peterse, J. L., and van 't Veer, L. J. (2005). Breast cancer metastasis: markers and models. *Nature Reviews Cancer*, 5(8):591–602. (Cited on pages 6 and 8.)

[Weigelt et al., 2012] Weigelt, B., Pusztai, L., Ashworth, A., and Reis-Filho, J. S. (2012). Challenges translating breast cancer gene signatures into the clinic. *Nature reviews Clinical oncology*, 9(1):58–64. (Cited on page 37.)

[Weinberg, 2007] Weinberg, R. A. (2007). Is metastasis predetermined? *Molecular oncology*, 1(3):263–264. (Cited on page 36.)

[Weinstein, 2002] Weinstein, I. B. (2002). Addiction to oncogenes–the achilles heal of cancer. *Science*, 297(5578):63–64. (Cited on page 45.)

[Winter, 1987] Winter, P. (1987). Steiner problem in networks: a survey. *Networks*, 17(2). (Cited on page 65.)

[Wirapati et al., 2008] Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B., Desmedt, C., Ignatiadis, M., Sengstag, T., Schutz, F., et al. (2008). Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res*, 10(4):R65. (Cited on page 55.)

[Wittekind and Neid, 2005] Wittekind, C. and Neid, M. (2005). Cancer invasion and metastasis. *Oncology*, 69(Suppl. 1):14–16. (Cited on page 9.)

[Wolpert, 1992] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259. (Cited on pages 77 and 78.)

[Wu et al., 2008] Wu, J. M., Fackler, M. J., Halushka, M. K., Molavi, D. W., Taylor, M. E., Teo, W. W., Griffin, C., Fetting, J., Davidson, N. E., De Marzo, A. M., Hicks, J. L., Chitale, D., Ladanyi, M., Sukumar, S., and Argani, P. (2008). Heterogeneity of breast cancer metastases: comparison of therapeutic target expression and promoter methylation between primary tumors and their multifocal metastases. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 14(7):1938–1946. (Cited on page 57.)

[Yang et al., 2014] Yang, R., Bai, Y., Qin, Z., and Yu, T. (2014). EgoNet: identification of human disease ego-network modules. *BMC Genomics*, 15:314. (Cited on page 71.)

[Yeung et al., 2001] Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987. (Cited on pages 20 and 107.)

[Zhang et al., 2005] Zhang, B., Kirov, S., and Snoddy, J. (2005). Webgestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic acids research*, 33(suppl 2):W741–W748. (Cited on page 51.)

[Zhe et al., 2011] Zhe, X., Cher, M. L., and Bonfil, R. D. (2011). Circulating tumor cells: finding the needle in the haystack. *Am J Cancer Res*, 1(6):740–751. (Cited on page 9.)

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320. (Cited on pages 26 and 94.)

Index