Unraveling Biological Information from Gene Expression Data using Advanced Clustering Techniques

Itai Kela

M.Sc Thesis submitted to the Feinberg Graduate School
Weizmann Institute of Science

Research conducted under the supervision of
Prof. Eytan Domany and Prof. David Givol

December 2001

Acknowledgments

I would like to take this opportunity to thank Prof Eytan Domany for his endless dedication and effort in guiding me throughout the year. I enjoyed working with him very much, and learned a lot from his experience.
I would also like to thank Prof David Givol who was very helpful and always found the time to share with me his outstanding knowledge.
I want to thank Gaddy Getz and Erel Levine for their endless patience and assistance.
I could not do it without you. Many thanks to Noam, Yariv, Uriel, Liat and Hilah for helping me in many ways.
Dr Irit Sagi has walked me through my first steps in research, and has equipped me with my fundamental scientific skills. For this I would like to thank her a lot. Many thanks goes also to Irit's group members, Arnon Henn and Oded Kleifeld.
Finally, I wish to thank my loving family for their support and for always being there for me.

Contents

Chapter 1

Abstract

My study focused on the use of new clustering methods to analyze large sets of gene expression data, obtained from experimental cell line in culture and from breast cancer clinical samples.

In my first study I used the temperature sensitive p53, expressed in the lung cancer cell line H1299, to analyze p53-regulated genes, utilizing oligonucleotide microarrays containing 7070 probes. Cycloheximide (CHX) was employed to inhibit protein synthesis in an effort to limit the p53-regulated genes to primary targets, by exclusion of possible secondary effects induced by newly synthesized proteins. By measuring gene expression at four or five time points I clearly showed that a group of genes exhibit consistent p53-dependent regulation, in either presence or absence of CHX. This group consists of less than 20% of the genes regulated by p53 in the absence of CHX, and was defined as primary targets for p53. The criteria used to group these genes (2.5 fold change in at least 3 time points) was somewhat arbitrary, and may have discriminated against genes which showed consistent changes but did not reach the 2.5 fold change at all the time points analyzed. I therefore used a clustering method, which allows the grouping of genes based on the kinetics of their expression, and showed that most of the primary genes group together in one or two clusters. These clusters indeed contained additional genes, with similar kinetic behaviors in their change of expression even though their levels of expression, did not reach the stringent criteria mentioned above. This group contains some of the established targets of p53 (e.g. Gadd45 and PCNA) and demonstrates the usefulness of this clustering method, since it overcame small experimental variations in measurement of hybridization intensity and relies on the pattern of expression in the presence or absence of CHX.

In the second part, I re-analyze the data that first appeared in the paper "Molecular portraits of human breast cancer" (Perou et al) [42]. The main goal of this paper was to develop a system for classifying tumors on the basis of their gene expression patterns. This paper characterizes gene expression profiles of 65 tumors and 19 cell lines, using cDNA microarrays, representing 8,102 human genes. Perou et al selected the subset of genes whose expression varied by at least 4-fold from the median over the samples, in at least three of the samples tested. This filtering process left 1753 genes, each of which is represented by 84 expression values. Since it was not possible to obtain meaningful partitions of the samples on the basis of

all these genes, Perou et al used their biological insight to generate a list of 496 “intrinsic genes” that were used to classify the tumors. This list contains those genes from the 1753 list that showed significantly greater variation in expression *between different* tumors than *between paired* samples from the same tumor. In my work I used the SPC method [12-13] and the CTWC algorithm [14] to analyze the data of Perou et al, aiming to compare our method of data analysis with that of Perou et al in order to see whether more insights can be obtained from the data collected by DNA microarrays. This is relevant to the work tested here (Perou et al) and the entire collection of expression data present in the literature.

I posed the following questions:

1. Do our methods of analysis reproduce the results obtained by Perou et al?

2 .Can I make observations that seem to be of interest and were not reported by Perou et al due to different clustering approaches?

The comparison revealed two major points; first, that the main findings of Perou et al shown in Fig. 5.3 can be found directly (and improved upon), starting from the entire set of 1753 genes and using the CTWC, without filtering the genes further (to the “intrinsic set” of 496 genes). Second, I find new tumor classifications that were not mentioned by Perou et al and group tumor samples according to new subsets of expression genes..

Chapter 2

DNA Chips: Methods for large scale monitoring of gene expression

Monitoring gene expression is important in many fields of biological research, since changes in the physiology of an organism or a cell will be accompanied by changes in the pattern of gene expression. The collection of genes that are expressed or transcribed from genomic DNA, sometimes referred to as the expression profile or the 'transcriptome', is a major determinant of cellular phenotype and function. The transcription of genomic DNA to produce mRNA is the first step in the process of protein synthesis, and differences in gene expression are responsible for both morphological and phenotypic differences as well as indicative of cellular responses to environmental stimuli and perturbations. Unlike the genome, the transcriptome is highly dynamic and changes rapidly and dramatically in response to perturbations or during normal cellular events such as DNA replication and cell division [1-2]. Knowing the extent and cause of a gene's expression is central to understanding the activity and biological roles of its encoded protein. In addition, changes in the multi-gene patterns of expression can provide clues about regulatory mechanisms and broader cellular functions and biochemical pathways. In the context of human health and treatment, the knowledge gained from these types of measurements can help determine the causes and consequences of diseases, how drugs and drug candidates work in cells and organisms, and what gene products might have therapeutic uses themselves or may be appropriate targets for therapeutic intervention. Several techniques for the analysis of gene expression at the mRNA-level are available. Methods such as *Northern blots* have the disadvantage of being inherently serial, involving measuring a single mRNA at a time, and of being difficult to automate. *Differential Display* [3] of amplified subsets of RNAs on a sequencing gel allows a broad search for expression difference, but the results generally are not quantitative, false positive are common, and characterization of positives requires additional cloning and sequencing. *Sequencing of cDNA* libraries is a more direct approach, but requires a great deal of sequencing and is not sensitive to the presence of less abundant mRNAs. The *Serial Analysis of Gene Expression* (SAGE) [4] is a clever and efficient variation of the cDNA sequencing approach. The method involves fairly complicated sample

preparation procedures, still requires a large amount of sequencing, and tends to be laborious and not particularly sensitive. Every one of these methods has some disadvantage, which renders them unsuitable if large numbers of expression products have to be analysed simultaneously.

The use of microarrays to monitor gene expression on large scale has recently received a great deal of attention [5-10]. Tens of thousands of transcript species can be detected and quantified simultaneously. This technology is based on hybridization of labeled RNA or DNA in solution ("target") to DNA molecules attached at specific locations on a high-density array (the "probes") [11]. The hybridization of a sample to an array is, in effect, a highly parallel search by each molecule for a matching or complementary partner on an 'affinity matrix'. In a simplified picture, the mRNAs (transcripts) are extracted from the cell, fluorescence labeled and hybridized to the array, where each transcript will quantitatively hybridize to its complementary target sequence. The fluorescence at each spot on the array is a quantitative measure corresponding to the expression level of the particular gene. Therefore, the major advance of DNA microarray technology, as compared to conventional techniques, is the capability of parallel screening of thousands of genes simultaneously. During recent years, DNA microarray technology has been advancing rapidly. The development of more powerful robots for arraying, new surface technology for glass slides, and new labeling protocols and dyes, together with increasing genome-sequence information for different organisms, including human, will enable us to extend the quality and complexity of microarray experiments [9]. Although many different microarray systems have been developed by academic groups and commercial suppliers, the most commonly used systems today can be divided into two groups, according to the array material: complementary DNA (cDNA) and oligonucleotide microarrays (Fig. 2.1).

2.1 cDNA microarrays

Array preparation. Probes for cDNA arrays (double strand DNA at average size of 1000 mer) are usually products of a polymerase chain reaction (PCR). Each probe represents a gene, and can be generated from the gene's cDNA clone. In simple terms, cDNA clones are made by extracting the mRNAs of interest from cells and then making a *complementary DNA (cDNA)* copy of each mRNA molecule present. This reaction is catalyzed by the reverse transcriptase enzyme of retroviruses, which synthesizes a DNA chain on an RNA template. The *DNA polymerase* enzyme converts the single-strand DNA into double-strand DNA molecules, and these are inserted into a plasmid or virus vector and cloned. Each clone

obtained in this way is called a cDNA clone, and the entire collection of clones derived from one mRNA preparation constitutes a cDNA library. The identification and extraction of a gene of interest from the cDNA library is done by a gene specific primer; radioactive or chemically labeled DNA probe, containing part of the sequence of the gene. Finally, the number of the gene's copies is amplified by PCR to generate the source of the cDNA sample that will be deposited on the chip. A micro sample (approximately one nanogram of cDNA material, or ~10^9 copies of 1kb cDNAs) of each cDNA is deposited and bonded on a glass surface at intervals of 100-300 μm, with each gene occupying a unique location (Fig. 1). Using this technique, arrays consisting of more than 30,000 cDNAs can be fitted onto the surface of a conventional microscope slide. Spotted cDNA arrays, allow a greater degree of flexibility in the choice of arrayed elements, particularly for the preparation of smaller, customized arrays for specific investigations. As a result, cDNA arrays have so far been the technique most frequently used in academic labs.

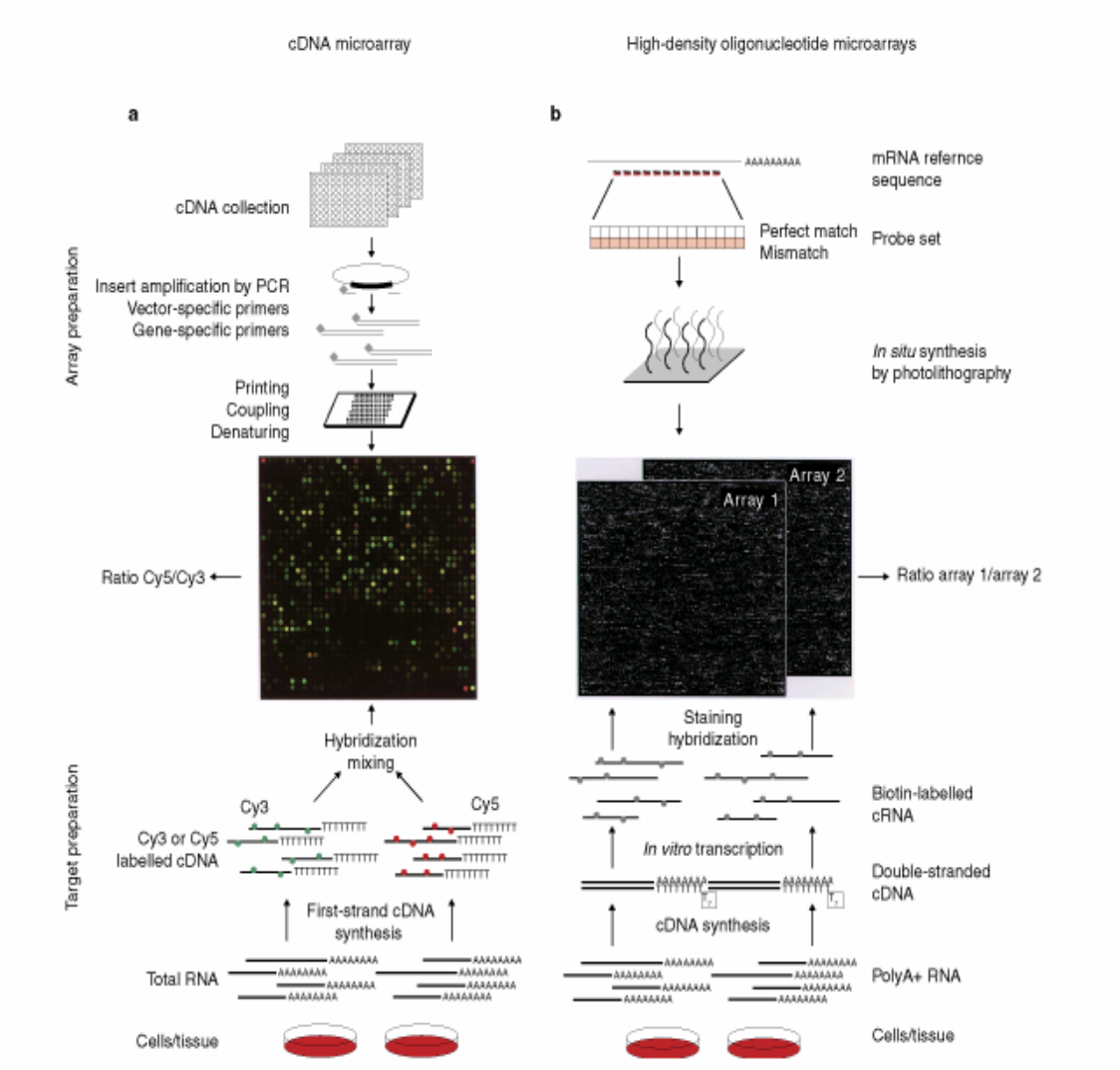

Figure 2.1. Schematic overview of probe array and target preparation for spotted cDNA microarrays and high-density oligonucleotide microarrays.

Target preparation. Messenger RNA (mRNA) molecules are extracted from normal or unaffected sample or "control" (e.g. Chromosome 6 suppressed cells in Fig. 2.2) and are reverse transcribed, by reverse transcription enzyme, to generate cDNA probes. During the reverse transcription, fluorescent-labeled nucleotides are incorporated into the synthesis of the cDNA, generating labeled cDNAs. Different mRNAs are extracted from another sample - the "experiment" (e.g. the Tumorigenic cells in Fig. 2.2), typically, these are the affected cells: exposed to a drug or toxic substance, taken from a tumor, or removed at a different time. The fluorescent labeling step is repeated to generate a second cDNA probe using a fluorescent molecule of different color (Fig. 2.2). Generally, green label is used for the control and red for the experiment.

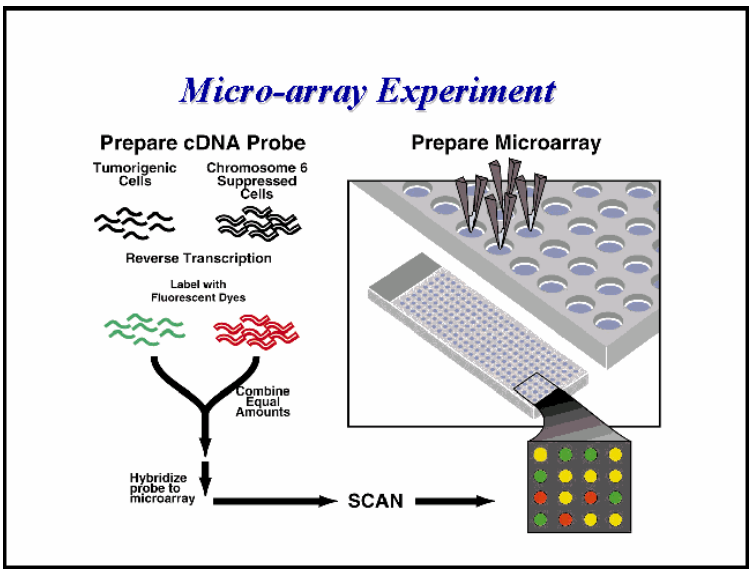

Figure 2.2. Two-color fluorescent hybridization for the analysis of gene expression. Messenger RNA (mRNA) molecules are extracted from normal or unaffected sample or "control" (e.g. Chromosome 6 suppressed cells in Fig. 2.2) and are reverse transcribed, by reverse transcription enzyme, to generate labeled cDNA probes. Generally, green label is used for the control and red for the experiment.

Hybridization. In the hybridization step the two fluorescent target samples are applied simultaneously to a single microarray, where they react competitively with the arrayed cDNA molecules. Following incubation (typically overnight), the microarray is washed off those probe molecules that did not find their cDNA counterpart. Genes that are highly expressed in the experiment's cells will have higher concentration of their RNA copies than those of the control; therefore, the hybridization level of these probes cDNA, labeled with Cye5, will be higher compared to the control. This results in domination of the red color on the corresponding spot (Fig. 2.2, 2.3). Genes with low expression level in the experiment will have low amounts of RNA compared to the control, resulting in domination of the corresponding spot by green color.

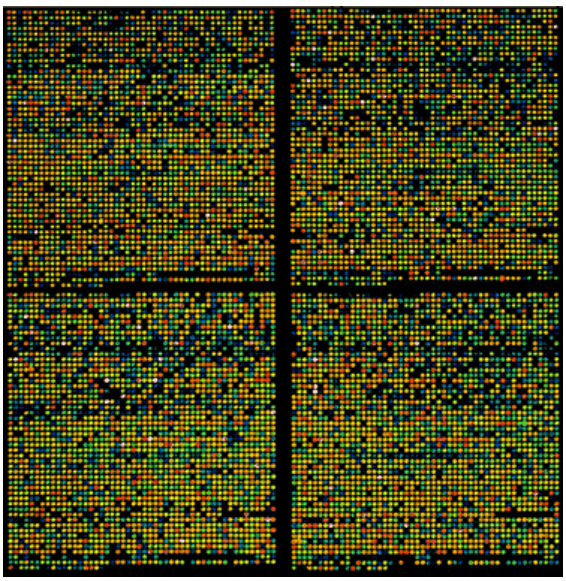

Figure 2.3. cDNA chip. The red color (Cye3) represents the control, the green color represents the experiment and the dark color represents no detection. The chip is scanned and the fluorescent intensity is determined at each spot.

After hybridization each element of the chip is scanned for both fluorescent colors. Because the sequence of the cDNA at each physical location is generally known or can be determined, and because the recognition rules that govern hybridization are well understood, the signal intensity at each position gives not only a measure of the number of bound molecules, which reflect the expression level of the gene, but also the likely identity of the molecules. The scanning operation is repeated for the second fluorescent label. The ratio of the two fluorescent intensities provides a highly accurate and quantitative measurement of the relative gene expression level in the two cell samples. When a microarray element shows no color, it indicates that the gene in that element was not expressed in either cell sample, and if a single color (green or red) appears, it indicates that the gene was expressed only in the corresponding sample, green for the control sample or red for the experiment. The appearance of both colors indicates that the gene was expressed in both cell samples.

Calculating gene expression. Each gene is represented by one element (spot) bounded to the array. After the scanning operation the *average intensity* and the *background* of each element, denoted by CH1I (CH2I), CH1B (CH2B), respectively, are measured (Table 2.1, Fig. 2.4). This is done, separately, for the two colors (Cye3 and Cye5). The final intensity of each element, corresponding to the two colors, is calculated as the follow:

Cye3 : $CH1D = CH1I - CH1B$, Cye5 : $CH2D = CH2I - CH2B$

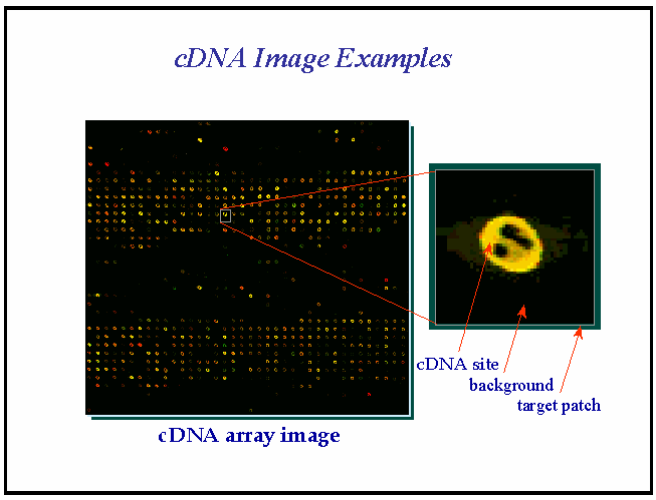

Figure 2.4. At each spot, the difference of the intensity averages between the cDNA site and its background is calculated.

The "final number", i.e. the value that we use through our analysis is calculated as the follow:

$$RAT2 = CH2D / CH1D$$

This value (RAT2, highlighted in red in Table 2.1) directly indicates the relative expression level of the gene. For example, an experiment that contains measurements done on 10 chips generates a data matrix (table) with 10 columns, corresponding to the 10 RAT2 columns, one from each chip, while the number of rows equals the number of the transcripts (genes) that were spotted on the chip.

HEADER	data
EXP	Name of the experiment based on the array id.
NAME	Name of the gene or sample that is at a given position.
TYPE	Indicates whether an element is a control or cDNA.
ACC	The GenBank accession number for an EST generated from the printed cDNA clone (This is the most stable handle for retrieving up to date information about this cDNA clone)
CLUS	The UniGene cluster assignment of the EST. This information is not stable and should be disregarded (see above)
CH1I	The average total signal for each spot for Cy3
CH1B	The local background calculated for each spot for Cy3
CH1D	The difference between the total and background for each element.
CH2I	The average total signal for each spot for Cy5.
CH2B	The local background calculated for each spot for Cy5
CH2D	The difference between the total and background for each element.
CH2IN	The average total signal for each spot for Cy5 normalized so that the average log ratio of all well measured spots is equal to 0.
CH2BN	The local background calculated for each spot for Cy5 normalized so that the average log ratio of all well measured spots is equal to 0.
CH2DN	The difference between the total and background for each element normalized so that the average log ratio of all well measured spots is equal to 0.
RAT1	CH1D/CH2D
RAT2	CH2D/CH1D
RAT1N	CH1D/CH2DN
RAT2N	CH2DN/CH1D
CRT1	and CRT2 Unused.
CORR	The correlation of the signal at each pixel of a spot in CH1 to CH2.
REGR	The ratio estimated by the slope of the line fit to the distribution of pixels in each spot.
RFLAG	Unused.
PLAT	The plate ID from which a sample was printed.

This is an example of the data:

EXP	NAME	TYPE	ACC	CH1I	CH1B	CH1D	CH2I	CH2B	CH2D	CH2IN	CH2BN	CH2DN	RAT1	RAT2	RAT1N
414	ZINC-ALPHA-2-GLYCOPROTEIN PRECURSOR		cDNA	W94583	763	461	302	920	447	473			348
414	ZINC FINGER PROTEIN IA-1		cDNA	R38640	605	501	104	567	457	110	442	356	86
414	ZINC FINGER PROTEIN HF.12		cDNA	W92135	845	508	337	1064	644	420	830	502	328
414	ZINC FINGER PROTEIN CLONE 647		cDNA	W86912	605	500	105	580	483	97	452	376	76
414	ZINC FINGER PROTEIN 76		cDNA	AA047454		773	580	193	914	657	257	713	512
414	ZINC FINGER PROTEIN 40		cDNA	N67788	854	494	360	994	571	423	775	445	330
414	"ZAKI-4 mRNA in human skin fibroblast, complete cds "		cDNA	W69771	1971	484	1487	2714	478	2236			
414	TRANSFORMING PROTEIN RHOC		cDNA	AA025344		1399	476	923	1764	508	1256	1376	396
414	TRANSFORMING PROTEIN RHOA		cDNA	AA047765		9244	559	8685	10629	548	10081	8294	427
414	TRANSFORMING GROWTH FACTOR BETA-1 BINDING PROTEIN PRECURSOR		cDNA	AA037699						2813	567	2246	3421
414	TRANSFORMING GROWTH FACTOR BETA 3 PRECURSOR		cDNA	AA040617				748	541	207	872	664	208
414	TRANSFORMING GROWTH FACTOR BETA 2 PRECURSOR		cDNA	N48082	653	513	140	668	504	164	521		
414	TRANSFORMING GROWTH FACTOR BETA 1 PRECURSOR		cDNA	R76211	596	532	64	596	509	87	465		
414	TRANSFORMATION-SENSITIVE PROTEIN IEF SSP 3521		cDNA	T47637	1374	567	807	1520	504	1016	1186		
414	TRANSFERRIN RECEPTOR PROTEIN		cDNA	AA055468		4026	542	3484	3937	568	3369	3072	443
414	TRANSDUCIN-LIKE ENHANCER PROTEIN 4		cDNA	AA035418			4964	620	4344	6995	806	6189	5458
414	TRANSCRIPTIONAL REPRESSOR PROTEIN YY1		cDNA	W31532	1009	545	464	1114	532	582	869	415	
414	TRANSCRIPTIONAL REGULATOR ISGF3 GAMMA SUBUNIT		cDNA	N89796	2096	561	1535	2469	548	1921	1926		
414	TRANSCRIPTIONAL ENHANCER FACTOR TEF-1		cDNA	H96798	4317	592	3725	4790	556	4234	3738	433	
414	TRANSCRIPTION INITIATION FACTOR TFIID 250 KD SUBUNIT		cDNA	R83300	997	590	407	1085	476	609			
414	TRANSCRIPTION INITIATION FACTOR TFIID		cDNA	N50549	1088	576	512	1163	595	568	907	464	
414	"TRANSCRIPTION INITIATION FACTOR IIF, BETA SUBUNIT "		cDNA	N73082	1064	575	489	1070	482	588			
414	"TRANSCRIPTION INITIATION FACTOR IIF, ALPHA SUBUNIT "		cDNA	W45575	1100	559	541	1063	492	571			
414	"TRANSCRIPTION INITIATION FACTOR IIE, BETA SUBUNIT "		cDNA	W94196	1125	609	516	1050	473	577			

Table 2.1. The lower table represents an example of a standard output file of cDNA micro array. The RAT2 column (marked by red circle) contains the numbers that we are using in our analysis. The top table describes each column of the output file; the important value is marked is the RAT2 value (marked in red).

2.2 Oligonucleotide - Arrays

Array preparation. For this technique the array is made by synthesis *in situ*, of short oligonucleotide (single strand, generally 18-25 mer), deposited either by photolithography onto silicon wafers (high-density-oligonucleotide array from Affymetrix, http://www.affymetrix.com) or by ink-jet technology (developed by Rosetta Inpharmatics, http://www.rii.com). The advantage of this method is that because sequence information alone is sufficient to generate the DNA to be arrayed, no time-consuming handling of cDNA resources is required. Also, probes can be designed to represent the most unique and specific parts of a given transcript, making the simultaneous detection of closely related genes or splice variants possible.

The oligonucleotide (oligo) array contains collections of approximately 20 pairs of probes for each of the RNAs being monitored. Each probe pair consists of two patches. One contains copies of a selected oligonucleotide (usually 20 to 25 nucleotides in length) that is perfectly complementary (refered to as a Perfect Match, PM) to a subsequence of a particular RNA. The second, companion patch contains identical oligonucleotides, except for a single base difference in a central position (referred to as a Mis Match on Fig. 2.5). The mismatch (MM) probe of each pair serves as an internal control for hybridization specificity. The analysis of PM/MM pairs allows low-intensity hybridization patterns from rare RNAs to be sensitively and accurately recognized in the presence of cross-hybridization signals. Hence, each gene is

represented by, usually 20 pairs of PM and MM of specific oligos, an opposed to the cDNA array, in while each gene is represented by copies of a single cDNA, deposited in one spot.

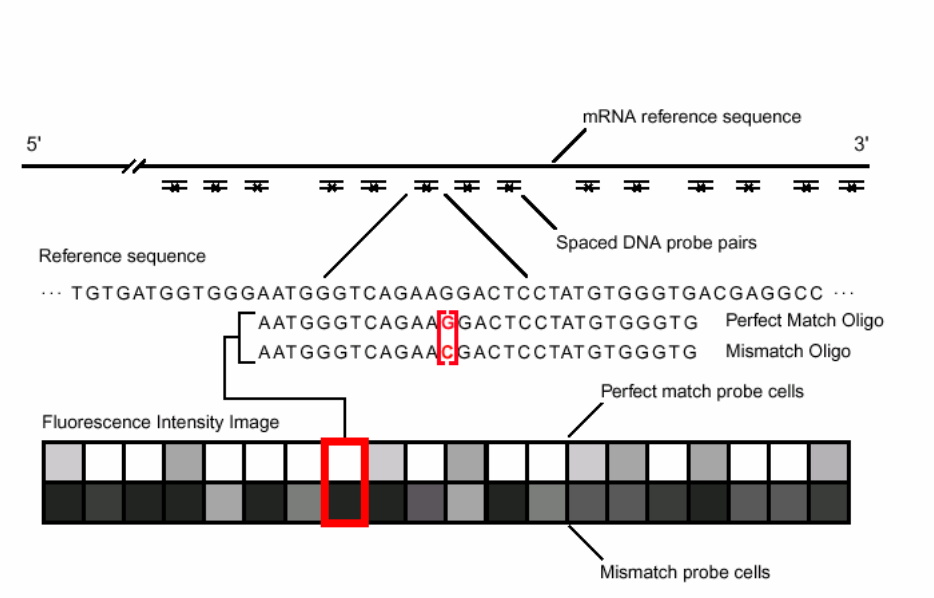

Figure 2.5. Oligonucleotide probes are chosen from the mRNA reference sequence, based on uniqueness criteria. For eukaryotic organisms, probes are chosen typically from the 3' end of the gene or transcript (nearer to the poly(A) tail) to reduce problems that may arise from the use of partially degraded mRNA. Each gene is represented by 20 PM and MM probe pairs. The use of PM minus MM differences, averaged across a set of probes, greatly reduces the contribution of background and cross-hybridization and increases the quantitative accuracy and reproducibility of the measurements.

Target preparation. Total mRNAs are extracted from different tissues or cell populations and pass hybridization reaction with dT primers (oligo of Thimidines) that contain a recognition site for T7 RNA polymerase (highlighted by a red circle on Fig. 2.6). The interesting mRNAs (transcripts) are the ones that are present in the cytoplasm; these transcripts are elongated with an Adenine chain (polyA), during the maturation process, to aids its exportation from the nucleus (the poly-A tail appears to have more functions as stabilization, and serving a recognition signal for the ribosome). For most genes, changes in polyA mRNA abundance are related to changes in protein abundance [11]. Extracting of the total RNA from the cells includes also the RNAs that are inside the nucleus. In order to "fish" only the polyA ones, dT primers are being in used. The dT primers contain a recognition site for T7 RNA polymerase are hybridized with the polyA mRNAs. After the primers hybridization process, a reverse transcription's reaction, carries out by a reverse transcriptase enzyme, leads the synthesis of the cDNA strand (Fig. 2.6). In the final step, a transcription reaction, carries out by the T7 RNA polymerase enzyme, take place, while biotin-labeled

nucleotides are incorporated into the synthesis cRNA molecules. This stage results in amplification of the RNAs molecules, approximately in 50 fold (up to 50 µg of labeled cRNA can be produced from 1 µg of mRNA) [9].

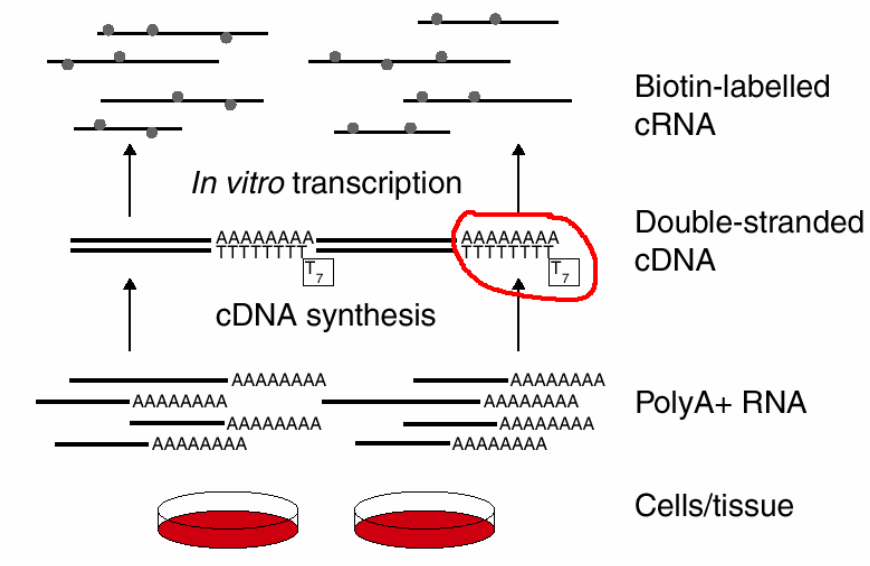

Figure 2.6. Target preparation for oligonucleotide arrays. A poly-A tail is added to the mature mRNA to aids its exportation from the nucleus. In order to "fish" only the poly-A ones, dT primers are being in used. The dT primers contain a recognition site for T7 RNA polymerase (highlighted in red) are hybridized with the polyA mRNAs. After the primers hybridization process, a reverse transcription's reaction, carries out by a reverse transcriptase enzyme, leads the synthesis of the cDNA strand. In the final step, a transcription reaction, carries out by the T7 RNA polymerase enzyme, take place, while biotin-labeled nucleotides are incorporated into the synthesis cRNA molecules.

The cRNA molecules, contain biotin-labeled nucleotides, are hybridized to the array; identify each kind of target sample is hybridized to a separate probe array. The fluorescent label procedure takes place through fluorescent label Avidin hybridization technique; the fluorescent Avidin molecules are bound to the Biotin's molecules, which are bound to some of the cRNA's nucleotides. In this step one Avidin binds to one Biotin molecule. In order to amplify the fluorescent signal another step of fluorescent labeling hybridization should be carry out; fluorescent labeled antibodies, where each antibody contains large number of fluorescent molecules, are added and hybridized with the Avidin molecules, results the amplification of the fluorescent signal.

A specially designed scanning microscope performs fluorescent imaging of the arrays. The entire array is read in less than 15 min, yielding a rapid and quantitative measure of each of the individual hybridization reaction.

The GeneChip software (Affymetrix, Santa Clara, USA) is the most commonly used tool to analyze the data generated from expression analysis probe arrays. Genechip algorithms are a

set of rules and calculations used to derive biologically meaningful results from hybridization intensities (see the GeneChip 3.1 Expression Analysis Algorithm Tutorial).

Calculating gene expression. 20 PM-MM pairs of specific oligos represent each gene. The *Average intensity* and the *background,* denoted by *A* and *B*, respectively, are measured for each of the 20 PM and the 20 MM cells (total of 40 cells). We denote by P_i (M_i) the difference between A_i and B_i of the corresponding PM (MM) cell:

$P_i = A_i^{(PM)} - B_i^{(PM)}$, $M_i = A_i^{(MM)} - B_i^{(MM)}$.

Calculating the *Average Difference* (*AvgDiff*) yields the "final number", i.e. the value that we use through our analysis. This value (highlighted in red in Table 2.2) directly indicates the expression level of the gene.

$$\text{AvgDiff} = \frac{1}{20}\sum_{i=1}^{20}(P_i - M_i)$$

Each experiment results in an Afymetrix output file (i.e. Table 2.2), where each gene is represented by several values. As I mentioned above, the value that we use through our analysis, denoted by *AvgDiff* , is the most important one.

I will present two studies, one of these analyzed data obtained from oligonucleotide arrays, and the other based on cDNA arrays.

Experiment: D:\Tesdata\Affymetrix\DGIVOLEXP1\11DGrc.EXP
Probe Array: Hu6800
Probe Array Lot: 99731
Operator Name: jasmine
Sample Type: David Givol 1st exp
Sample Description:
Project:
Comments:
Reagents:
Reagent Lot:

Algorithm Paramet 11DGrc
BF= SDT=102.6 SRT=1.5 ABS={(3.0 RawQ=27.62
QMult=2.80 RL=10.0 HZ=4 VZ=4 BG=2
CT=102.6 PCT=80 DD={(3.C AvgNCT=2 FCMin=1.00
FCMax=1.00 STP=3.0 TGT=12(NF=1.000(SF=0.928356
SFGene=All

Experiment: 11DGrc
Corner+ Avg:3083 Count:32
Corner- Avg:45203 Count:32
Edge+ Avg:592 Count:8
Edge- Avg:45097 Count:8
Background Avg: 862 Stdv: 38. Min: 811 Max: 940
Pixels Avg: 35 Stdv: 1.8 Min: 25 Max: 36

No	Experiment Name	Probe Set	Positive	Negative	Pairs	Pairs Used	Pairs InAvg	Pos Fraction	Log Avg	PM Excess	MM Excess	Pos/Neg	Avg Diff	Abs Call
1	1DGrc	\B000114_a	8	0	20	20	18	0.4	1.19	0	0	Undef	125	P
2	11DGrc	\B000115_a	8	1	20	20	17	0.4	1.38	0	0	8	177	P
3	11DGrc	\B000220_a	18	0	20	20	19	0.9	4.19	0	0	Undef	527	P
4	11DGrc	\B000409_a	12	1	20	20	18	0.6	2.53	0	0	12	307	P
5	11DGrc	\B000449_a	11	0	20	20	19	0.55	3	1	0	Undef	552	P
6	11DGrc	\B000450_a	12	1	20	20	19	0.6	2.24	0	0	12	513	P
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
7010	11DGrc	\B000905_a	9	0	20	20	20	0.45	1.67	0	0	Undef	182	P
7011	11DGrc	\B001106_a	19	0	20	20	19	0.95	5.14	1	0	Undef	1547	P
7012	11DGrc	\B001325_a	10	2	20	20	19	0.5	1.98	0	0	5	836	P
7013	11DGrc	\B002314_a	6	2	20	20	18	0.3	0.78	0	0	3	125	A
7014	11DGrc	\B002315_a	10	2	20	20	19	0.5	1.4	0	1	5	470	P
7015	11DGrc	\B002318_a	9	5	20	20	19	0.45	1.22	0	0	1.8	932	A

Table 2.2. An example of output file created by the GeneChip software. The data used in our analysis is located in the column named AvgDiff. This column represents the expression level of each transcript.

Chapter 3

Clustering methods

In this chapter the main data analysis tool used in this work, namely data-clustering, will be introduced. A clustering method, Super-Paramagnetic Clustering (SPC) [12-13], which was extensively used in this work, will be described in details. The data-mining method, Coupled Two-Way Clustering (CTWC) [14], used for analysis of breast cancer data in chapter 5, is also reviewed.

3.1 The Clustering Problem

The clustering problem can be formally defined as follows. Partition N given points into k groups (clusters) so that two points that belong to the same group are, in some sense, more similar than two that belong to different ones. In fact, the problem is inherently ill-posed, i.e. any given set of points can be clustered in drastically different ways. Difficulties and ambiguities may arise for a variety of reasons. For example, there may be data points that do not belong to either cloud; the shapes of the clouds may be complex and their density non-uniform. The most important source of ambiguity is that the manner in which data "should" be clustered depends on the desired resolution. What appears as a single cloud may turn out, when examined at higher resolution, to be composed of several sub-clusters.

Consider the example given in Fig. 3.1. At a very low resolution, seen for example by a man standing a few meters from the page, all eight points just sit together, and no structure can be observed. A closer look would identify two clusters – 4 points on the right and 4 on the left. At an extremely high resolution, seen from the point of view of an ant traveling on the paper, each point resides in its own cluster.

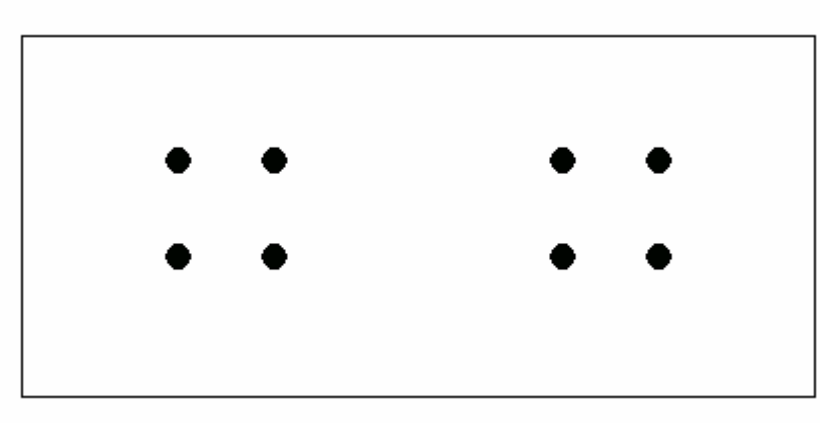

Figure 3.1. Eight points in two dimensional space.

3.2 The Super-Paramagnetic Clustering Algorithm (SPC)

The Super-Paramagnetic Clustering algorithm (SPC) [12-13] is a clustering method based on the physical understanding of disordered granular magnets. In what follows I first introduce a model for such physical systems, and review some of its physical properties. This is given in Sub-Section 3.2.1. In Sub-Section 3.2.2 I describe how these properties are used for data clustering. This is demonstrated in Sub-Section 3.2.3 on the toy problem of Fig. 3.1, where the physical analog can be solved exactly. In typical cases, however, exact solution cannot be obtained. Sub-Section 2.2.4 describes an approximate method used in these cases.

3.2.1 The physics behind SPC

SPC is based on a model for a magnetic system, which is known in the physics literature as the "Potts model". The model consists of small magnetic elements (spins) which interact with one another. This interaction can be mimicked by an energy function, which is lower when two spins are aligned and higher when they are not. Hence, at the minimum energy, all spins are aligned. The magnetic interaction decreases rapidly with the distance between the spins. The model assumes that two spins which are not neighbors (in some sense) experience no interaction between them.

A simple uniform magnet can be described as a lattice of spins (Fig. 3.2). The interaction between neighboring spins on this lattice is some constant, J>0, while the interactions between spins which are not neighbors is just zero. The energy function of such a magnet can be written as

(1) $$H = \sum_{\langle i,j \rangle} J\left(1 - \delta(s_i, s_j)\right).$$

Here s_i is the direction of the spin i, and the summation is over pairs of neighbors. It turns out that two very different temperature regimes exist for this magnet. At the lower temperature regime the spins in the system follow the energy bias and tend to be aligned. In the higher temperature regime the energy bias is ignored, and each spin points in any desirable direction.

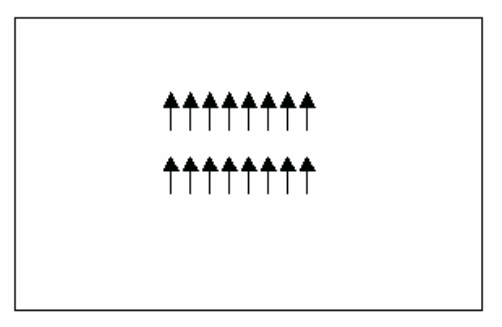

Figure 3.2. A model of a simple uniform magnet that can be described as a lattice of spins that tend to be align.

There are two mathematical measures, which enable one to distinguish the two regimes: a global measure and a local one. The global measure is called magnetization and is defines by $m = \frac{qN_{\max} - N}{(q-1)N}$, where q is the number of available directions for each spin, N is the number of spins, and N_{max} is the number of spins pointing in the most "popular" direction. In the ordered low-temperature regime $N_{max} \approx N$, so $m \approx 1$. In the high-temperature regime, $N_{max} \approx N/q$, so $m \approx 0$.

The local measure is the correlation between two spins. The correlation may be defined as the probability to find the two spins aligned. When properly defined, the correlation should be of the order of 1 in the ordered regime, where all spins tend to be aligned, and in the order of 1/q in the disordered regime. Of course, one expects the correlations to decrease smoothly from 1 to 1/q as the temperature is raised, with a sharp drop at the temperature separating the two regimes. This temperature is referred to as the transition temperature.

The idea behind SPC is to place a spin at every data point. In reality, the data points are not positioned on a uniform lattice (as in the simple magnet model) and therefore SPC uses a more general model for a **disordered granular magnet**. This material consists of isolated magnetized areas scattered randomly in a non-magnetic background. Such a magnet typically has strong interactions between spins, which reside in the same magnetic domain, and weak interaction between spins in different domains (Fig. 3.3). More precisely, the interaction between two neighboring spins decreases with the distance between them. Note that two spins may be considered neighbors even when they do not reside in the same domain. The interaction parameter J of the energy function (1) is replaced in the summation by J_{ij} , which is a function of the distance between spin i and spin j.

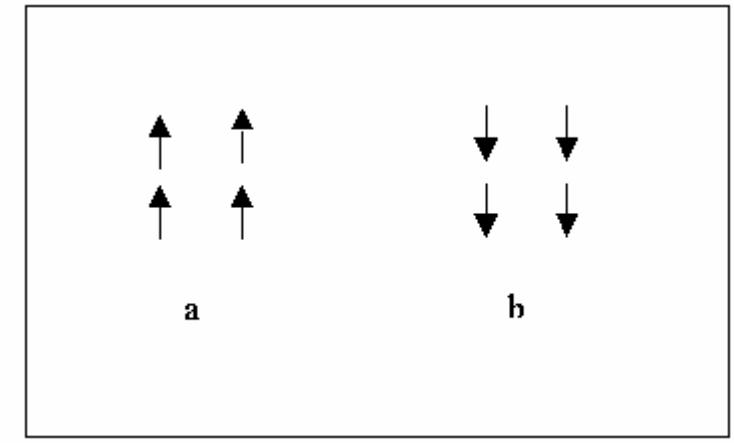

Figure 3.3. A model of disorder magnet, such a magnet typically has strong interactions between spins, which reside in the same magnetic domain, and weak interaction between spins in different domains

A granular magnet has three phases, depicted in Fig. 3.4: Ferromagnetic (low-temperature), super-paramagnetic (intermediate-temperature) and the paramagnetic (high-temperature) phases. Fig. 3.4 demonstrates a two dimensional example of a simple granular magnet. The magnet contains two domains, a and b. Each of these domains consists of 4 spins, denoted by arrows. At low temperatures (T), the system is in the **Ferromagnetic phase**: here the magnetization is 1, and the system shows high correlations between all pairs of spins. At higher temperatures, the system passes to the **Super-Paramagnetic phase**. In this phase a pair of spins inside a domain show high correlations as before, but pairs from different domains are uncorrelated. At very high temperatures, the system passes to the **Paramagnetic phase**, in which all pairs of spins are uncorrelated.

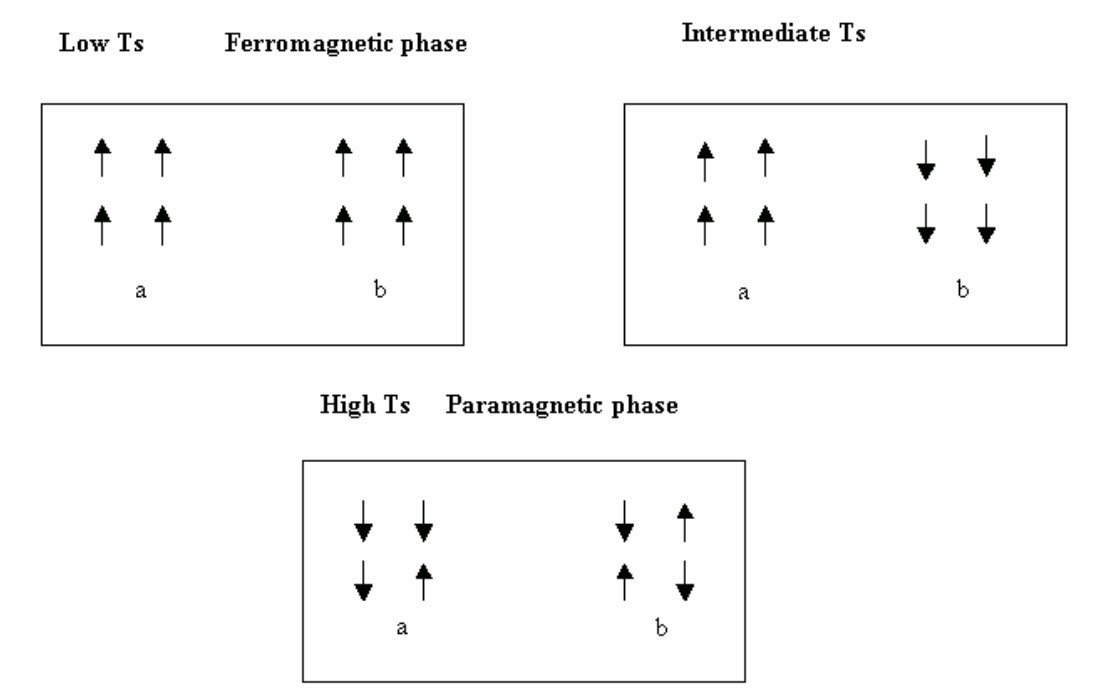

Figure 3.4. A two-dimensional example of a disordered magnet. At low Ts the system is in the Ferromagnet phase; the spins, denoted by arrows, inside each domain, a and b, and between the two domains, are correlated. At intermediate Ts the system passes to the Superparamagnetic phase; the spins inside each domain are correlated as before, but the two domains are uncorrelated. At high Ts the system is in the Paramagnetic phase; all the spins are uncorrelated.

3.2.2 How SPC applies the physics to cluster data

I now turn to describe how SPC applies the physics of disordered granular magnets to cluster data. SPC positions a spin at every data point. At a given temperature SPC assigns two points to the same cluster if the correlation between them exceeds ½. To calculate the correlation between the neighboring pairs, SPC uses a celebrated result from statistical mechanics: At a certain temperature T, the probability to find the system at a given configuration S is given by $P(S) = \frac{1}{Z}\exp\left(-\frac{H(S)}{T}\right)$, where H(S) is the energy of the configuration, and Z is the normalization factor. It is now straightforward to calculate any average, and in particular the correlation between spin i and j : $C_{ij} = 2\sum_{S\in\{S'\}} P(S) - 1$, where {S'} is the set of all configurations in which spin i and spin j point in the same direction. $C_{i,j}$ ranges between 0 and 1. This equation is a private case for q=2.

The temperature T controls the resolution in SPC. As the temperature is raised, domains, which are further apart, become uncorrelated. When going from the ordered ferromagnetic phase to the disordered paramagnetic phase, the system may go through several different super-paramagnetic phases, in each of which some domains may be correlated among themselves while uncorrelated with others.

3.2.3 Clustering toy data

I will now demonstrate the ideas behind SPC using a simple toy data set depicted in Fig. 3.1. The data set contains 8 points. Each data point is assigned a binary spin, i.e. a spin, which can assume one of two possible states (q=2). The interactions between the spins are represented in Fig. 3.5A by lines: dashed lines stand for weak interactions (J=1), while continuous lines stand for strong interactions (J=10). Note that the strengths of the interactions decrease with distance. The configurations Fig. 3.5B and Fig. 3.5C are examples of two (out of 256) possible configurations. In Fig. 3.5B is a 'low energy' (H=2) configuration with pairs (2,5) and (4,7) not aligned (neighboring spins are in different states). Fig. 3.5C depicts a 'high energy' (H=20) configuration in which pairs (8,6) and (8,7) are not aligned.

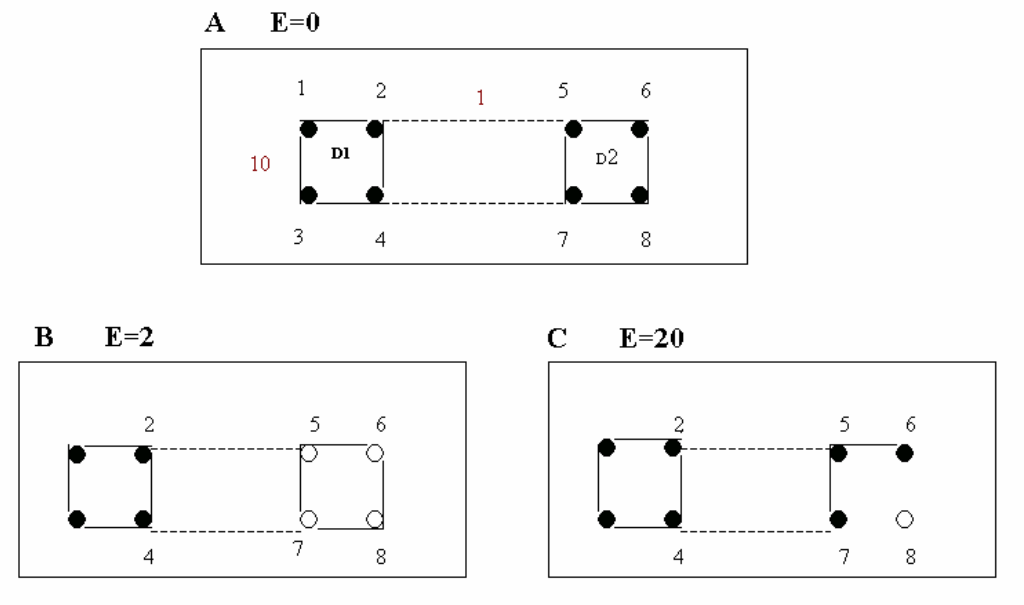

Figure 3.5. (A) One of the states of the system at *T*=0; all the points (spins) are in the same state. Inside each domain the points are connected by continuous lines, which stand for strong interactions (*J*=10). The two domains, D1 and D2, are connected with dashed lines; weak interactions (*J*=1). (B) A 'low energy' configuration. (C) A 'high energy' configuration.

We now proceed to cluster this data in the spirit of SPC. The differences between this demonstration and SPC are discussed in the last sub-section. In the toy model there are 256 (2^8) possible configurations. Each configuration is readily assigned an energy cost and a probability, as defined above. It is now straightforward to calculate the correlation between neighboring spins, as a function of temperature. These functions are shown in Fig. 3.8.

Fig. 3.6 represents the behavior of the $\log_{10}$ of P for each of the 256 states, as a function of T. Note that the energy is degenerate (there is more than one configuration at a given T). At low temperatures only the 'low energy' configurations (pointed by the upper arrow) get very high probabilities; log (P) = 0 (P=1). The 'high energy' configurations (pointed by the lower arrow) can be found with very low probabilities at these temperatures. As seen in the figure, as T increases the probabilities of all the configurations tend to be equal. At high temperatures all the interactions (weak and strong) have equal probability to be broken.

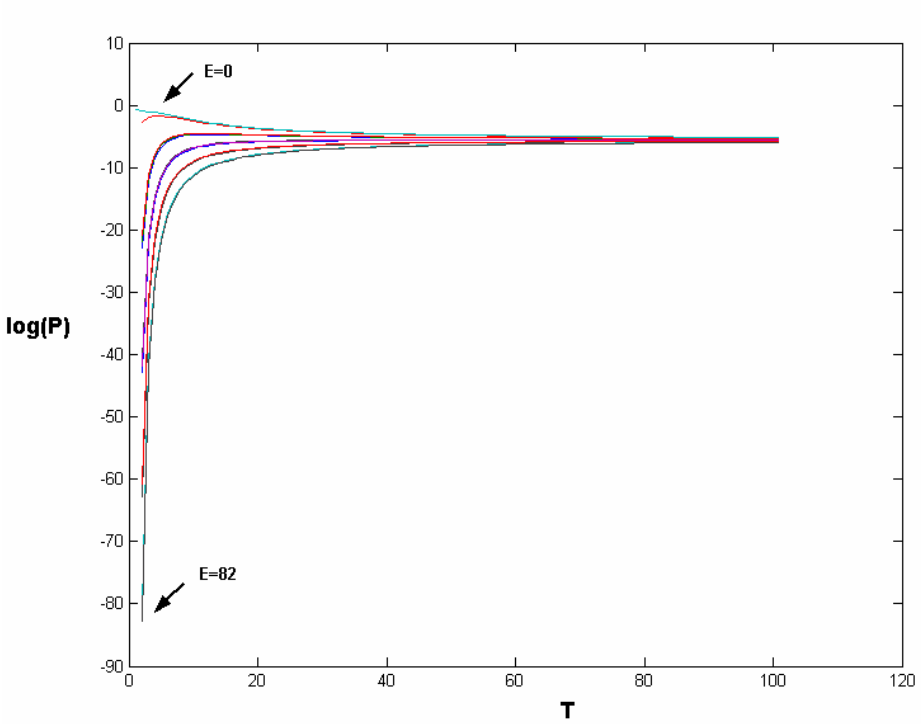

Figure 3.6. This figure represents the behavior of the log of P as a function of T for all the 256 microstates. At low temperatures, the 'low energy microstates (pointed by the upper side arrow) get very high probability (≈1), and the 'high energy' microstates' get very low probability (<<1). At high temperatures are equally likely.

Figure 3.7 represents the probability (P) as a function of T for the two configurations of Fig.5B and Fig. 3.5C. The blue line is for the 'low energy' configuration (shown in Fig. 3.5B). It reaches a high probability at low temperatures, and its probability decreases with a very sharp slope. The 'high energy' configuration (shown in Fig. 3.5C) is plotted with a red line has very low probability at low temperatures. As T increases, the difference between the probabilities of these configurations tends to 0.

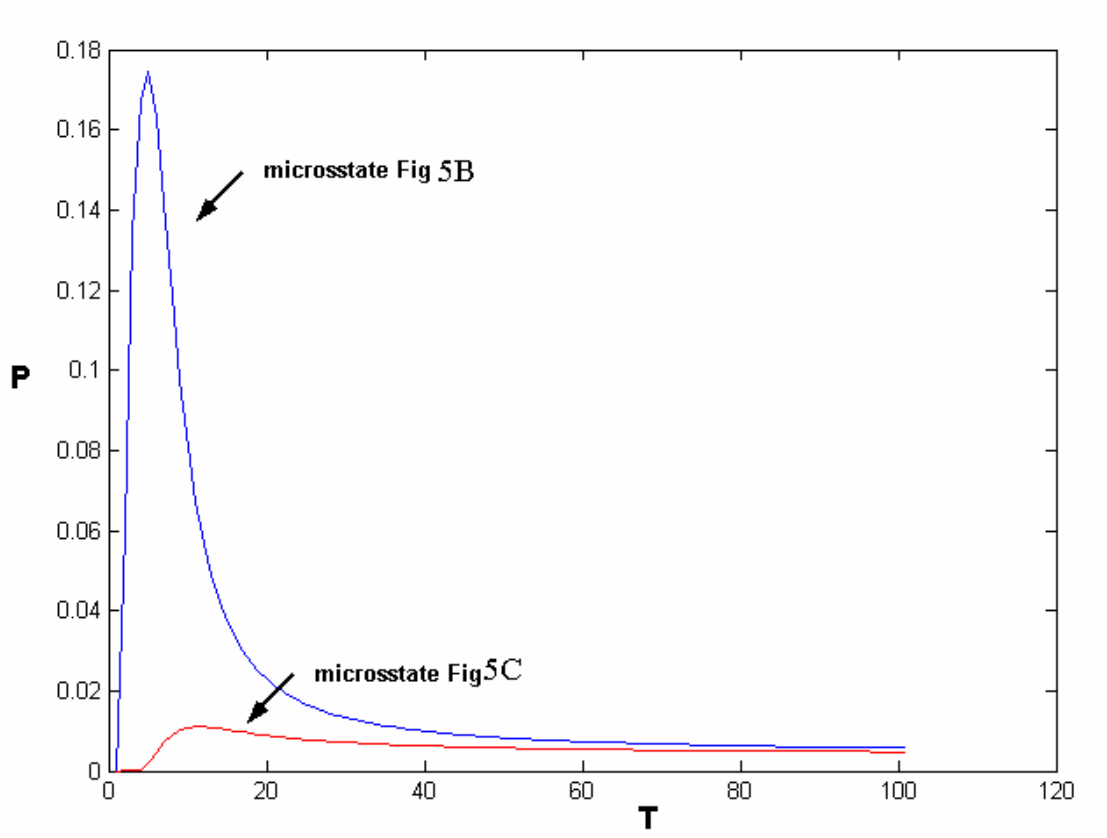

Figure 3.7. This figure represents the probability (P) as a function of T for the configurations that appear in Fig.5B and 5C. Blue - The 'low energy' configuration (Fig.5B) . Red - The 'high energy' configuration (Fig.5C).

Figure 3.8 concentrates on the correlation value of two pairs, (2,4) a strong interacting pair and (2,5) a weak interacting pair.

C(2,5) - The weak interacting pair J(2,5)=1. As shown in Fig. 3.8 (red line), indeed C(2,5) decreases at relatively low temperatures below 0.5; at T1=3 the spins 2 and 5 will not be assigned to the same cluster.

C(2,4) - The strong interacting pair J(2,4)=10. This bond is stronger than C(2,5). Looking at Fig. 3.8 one sees that, indeed, the C(2,4) decreases more slowly (blue line); it decreases below 0.5 at higher temperature T2=12.

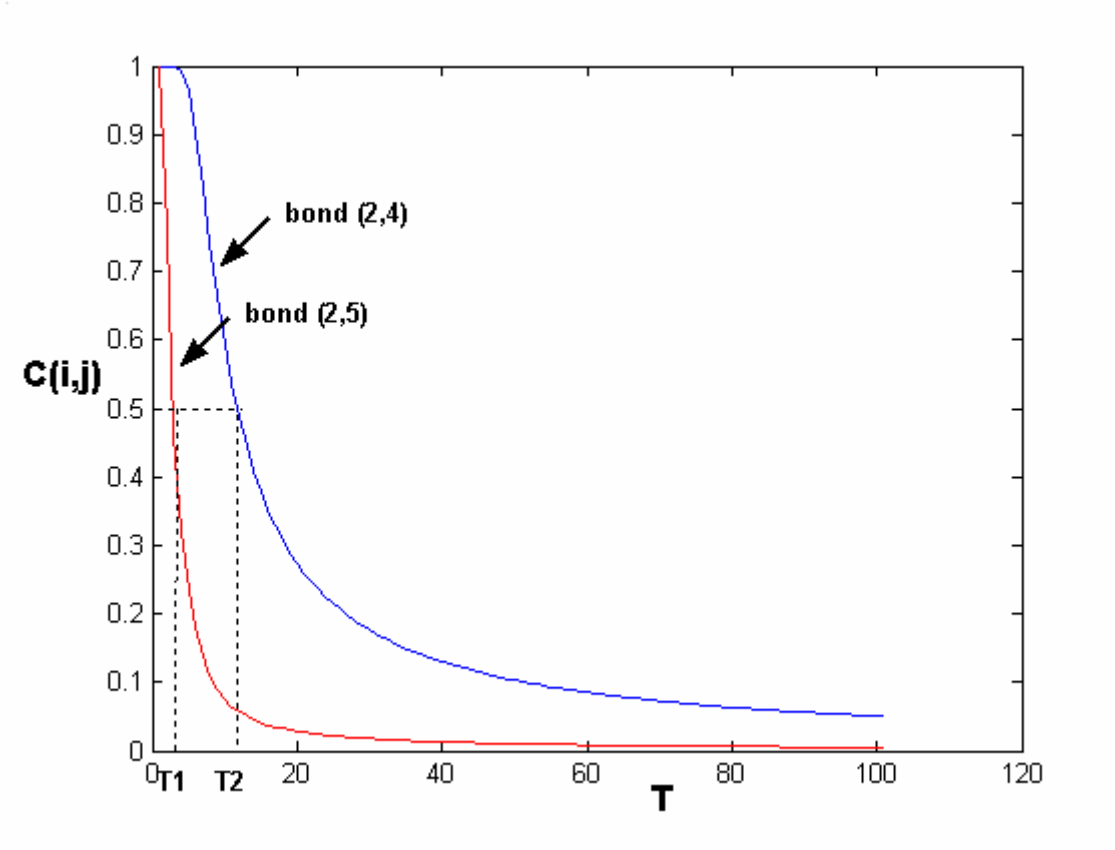

Figure 3.8. This figure demonstrates the behavior of *C*(2,5) and *C*(2,4) as a function of T. the correlation of the weaker interaction, *C*(2,5), decreases mach faster below 0.5 (red line) than *C*(2,4) (blue line); the stronger interaction. The temperatures T1 and T2 are the critical temperatures in which the pairs 2,5 and 2,4 are disconnected, respectively.

Figure 3.9 represents the range of temperatures of the three phases, ferromagnet, super-paramagnet and paramagnet. In my model I defined only two strengths of interactions, J=1 and J=10. When $0 \leq T \leq 3$ all the pairs with J=1 and J=10 are still connected, therefore the system is in the **ferromagnetic state** (yellow); at this range there is one big cluster that contains 8 points. When $3 < T \leq 12$ the system passes to the **super-paramagnetic phase** (red). The correlation, C, between the pairs with weak interaction (J=1) decrease below 0.5, therefore are disconnected. In this phase, two clusters are revealed; cluster A contains the points 1,2,3,4 and cluster B contains the points 5,6,7,8. When T>12, the system passes to the **paramagnetic phase** (purple). Above T=12 the correlations, C, between all pairs decrease

below 0.5. In this range of temperatures the system contains 8 clusters; each point is a separate cluster.

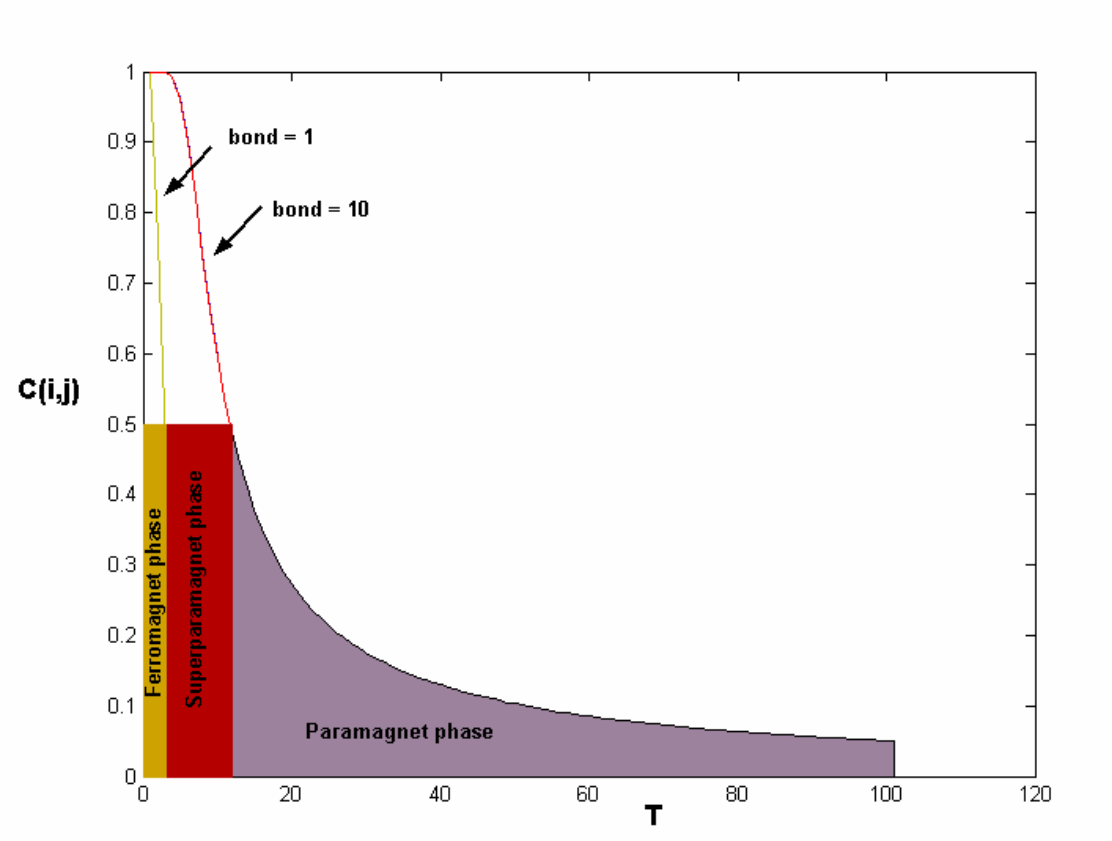

Figure 3.9. This figure represents the range of temperatures of the three phases: ferromagnetic ($0 \leq T \leq 3$), super-parmagnetic ($3 < T \leq 12$) and paramagnetic (T > 12).

3.3 The CTWC algorithm

The main idea of CTWC is to identify subsets of the genes and samples, such that when one of these is used to cluster the other, stable and significant partitions emerge. The CTWC process starts by clustering the data in two ways. The first is to look for clusters of genes that act correlatively on the different samples, considering the $N = n_g$ genes as the objects to be clustered, each represented by its expression profile, as measured over all the samples, that play the role of the features. Hence we cluster $N = n_g$ points in a $D = n_s$ dimensional space. The other way is to view the samples as the $N = n_s$ objects to be clustered, with the n_g genes' levels of expression playing the role of the features, representing each sample as a point in a $D = n_g$ dimensional space. Samples with similar or related expression profiles that may be due to similar cellular processes taking place will be grouped together.

The philosophy behind the next stages of CTWC is to narrow down both the features that are used and the data points that are clustered. Our assumption is that only a small subset of the genes participates in any cellular process of interest, which takes place only in a subset of the samples. By focusing on small subsets we lower the noise induced by the other samples and genes. We look for pairs of relatively small subsets of features, F_i (either genes or samples), and of objects O_i, (samples or genes), such that when the objects in O_i are clustered based on the features in F_i, stable and significant partitions are generated. The CTWC algorithm produces such (O_i, F_j) pairs in an iterative clustering process.

The CTWC can be used in conjunction with several clustering methods, but I present here only results that were obtained using the Super-Paramagnetic Clustering algorithm (SPC), which is especially suitable for gene microarray data analysis due to its robustness against noise, and for usage within CTWC because of its "natural" ability to identify stable clusters.

A variety of factors, such as cell type, cell phase, external signals and more, influence gene expression. Therefore the measured expression levels are the result of all these processes mixed together. The goal is to separate and identify these processes and to extract as much information as possible about them. It is most likely that only small subset of the genes present on the microarray plays a role in a particular process of interest. The large majority of these genes constitute a noisy background, which may mask the effect of the small relevant subset. The same may happen with respect to samples. The approach is to find pairs of subsets (O_i, F_i), which define particular submatrices that lead to "meaningful" clusters. CTWC

provides an efficient way to generate such pairs of object and feature subsets by an iterative process, that restricts the possible candidates for such subsets; we consider and test only those submatrices, whose rows (columns) belong to genes (samples) that were identified (in a previous iteration) as a stable cluster. The iterative process is initialized with the full matrix, i.e. the sets of all genes (G1) and of all samples (S1) are used as (both) features and objects, to perform standard two-way clustering. Denote by GI and SJ, with I,J = 2, 3,…, stable clusters of genes and samples found in this first step. Every pair (GI, SJ), made of clusters obtained so far, defines a submatrix of the expression data; for every such submatrix we perform two-way clustering. The resulting stable gene (or sample) clusters are also denoted by GI' (or SJ'). Each cluster is stored in one of two "registers of stable clusters"; gene clusters in register G and samples in S.

Chapter 4

Identification of primary and secondary target genes regulated by p53, using DNA microarrays

4.1 Introduction

4.1.1 Gene expression and cell identity

Different cell types of a multicellular organism contain the same DNA sequence.

Many processes are common to all cells and therefore they have many proteins in common. For example, the major structural proteins of the cytoskeleton and of chromosomes as well as some of the proteins essential to the Endoplasmatic Reticulum and Golgi membranes or ribosomal proteins are expressed in all cells [15]. These families of genes that are expressed in all cells with homogeneous levels are known as the 'housekeeping genes'. Studies of the number of different mRNA sequences in a cell suggest that a typical higher eucaryotic cell synthesizes 10,000 to 20,000 different proteins. It is commonly assumed that only a small number (perhaps several hundreds) of proteins suffice to generate large diversity in cell morphology and behavior. Cell differentiation mostly depends on changes in gene expression. Different cellular signals influence the genes expression patterns, and lead to the generation of different cell families with different properties and functions. Cells are sensitive to external and intrinsic signals during all theirs life; *growth* signals promote cell proliferation, *stress* signals (external or intrinsic to the cell) can stop proliferation and drive the cell to a growth arrest phase. Some signals may drive the cells to an irreversible process, known as "programmed cell death" or "apoptosis". Hence, normal response of cells to diverse signals (i.e. growth, growth arrest, apoptosis) is very important to regulate and prevent abnormalities in cell's functions. Different defects in cell properties, like loss of sensitivity to stress signals or over sensitivity to growth signals, can lead to abnormalities and later on, transform cells to

be malignant. These abnormalities can occur as results of loss of function or gain of function of two central types of genes; 1) tumor suppressors genes, and 2) oncogenes, that play a central role in growth processes. Mutations in such genes might transform the cell to be a cancerous one. In both cases, the cells lose their growth control.

Gene expression is highly regulated; its control can occur at four main levels; (1) **transcriptional level**, controlling when and how often a given gene is transcribed (2) **RNA processing level**, controlling how the primary transcript is spliced or otherwise processed (3) **translation level**, controlling the rate at which mRNAs in the cytoplasm will be translated into protein (4) **degradation level**, regulating the stability of the protein. The transcription level is considered the most important regulation phase.

4.1.2 Activation of gene expression by transcription factors

Transcription of a gene begins when an RNA polymerase molecule binds to a promoter sequence. The promoter is a specific DNA sequence located in up-stream region to the open reading frame (the transcribe region of the gene) that directs RNA polymerase to bind to DNA, to open the double helix, and to begin synthesizing an RNA molecule. In eucaryotic cells this process is complex, and requires two groups of regulatory proteins; 1) **general** transcription factors, and 2) **specific** transcription factors. The first group contains a small set of genes, abundant in all cells, and participates in the transcription process in all genes. These factors assemble on the promoters in order to recruit the RNA polymerase to begin the transcription. The second group contains a large set of transcription factors, which show heterogonous expression levels according to the cell type. These factors bind to the regulatory region of a gene, to specific target sequences. This is a particular nucleotide sequences, each typically less than 20 nucleotide pairs in length, functional as a fundamental components of genetics switches by serving as a recognition sites for the binding of specific gene regulatory proteins. Some of the transcription factors lead to induction of transcription, some lead to repression. Some transcription factors may function both as an inducer or a repressor, depending on additional factors. The location of the regulatory sequences over the promoter can be diverse; some sites are located near the start point of the transcription; in contrast, there are some regulatory sites (mainly in supreme eucaryotic species), that are located far away from the start point; sometimes more than 2000 bases up-stream. Some regulatory sites can be located inside introns, down-stream to the transcription start point. The transcription

control is very complex, and the outcome (induction or repression), is a result of complex combinations between all the factors, that bind to the promoter in a specific event. Although some transcription factors may work individually, most acts as part of a complex of several polypeptides, each with a distinct function. This complex often assembles only in the presence of the appropriate DNA sequence. An individual transcription factor can often participate in more than one type of regulatory complex. A protein may function in one case as a part of a complex that activates transcription, and in a different case as a part of a repression complex. Hence complex combinations between diverse regulatory factors control which of the thousands of genes in a cell will be transcribed and expressed.

4.1.3 The tumor-suppressor protein - p53

The p53 gene and its protein product have become the center of intensive study since it became clear that approximately 50% of human cancers contain mutations in this gene. Most human cancers contain mutation in the p53 gene or a functional defect in the p53 pathway, highlighting its importance for preventing tumorigenesis. p53 is a tumor-suppressor gene that has been called the “cellular Gatekeeper” and the “Guardian of the Genome” for its well-documented activities in causing cell cycle arrest or apoptosis in response to a variety of DNA damage. Although not required for normal development, p53 is a critical player in the prevention of tumor development.

Normally, the amount of p53 protein in the cell is kept at a low level because its relatively short half-life. Variety of conditions, as represented in Fig. 4.1, can lead to rapid induction of p53 activity. The common denominator of these conditions is their relation to stress. Such conditions include DNA damage as well as damage to components involved in the proper handling and segregation of the cellular genetic material. The rapid induction of p53 in response to genomic damage serves to ensure that cells carrying such damage are effectively

taken care of.

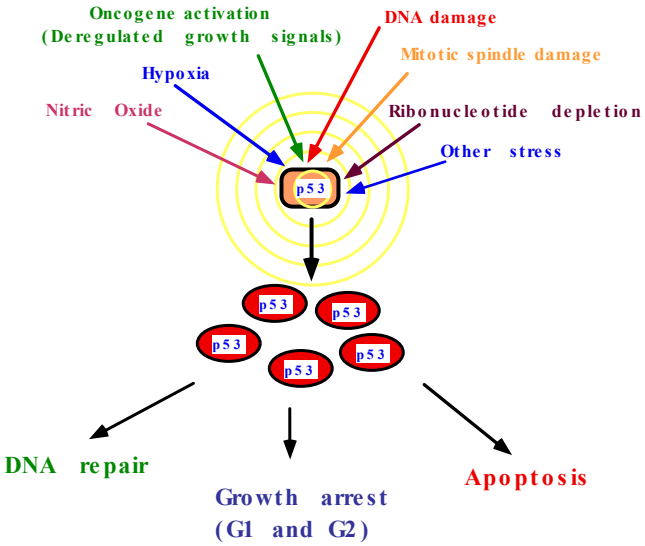

Figure 4.1. Various stress signals that lead to the activation of the tumor suppressor p53.

4.1.4 P53 as a transcription factor

The function of p53 as a tumor suppressor is mainly due to its activity as a transcription factor that activates many genes in response to various types of stress [16]. This may be the basis for p53 protection of cells against DNA damage and various stress conditions that lead usually to growth arrest or apoptosis [17]. p53 is a sequence-specific transcription factor; it binds to specific sequences within the cellular DNA, and activates the transcription of genes that contain such sequences in their promoters and regulatory sequences. The consensus binding site for p53 is defined as two copies of the 10-bp motif, 5',PuPuPuC(A/T)(T/A)GPyPyPy-3', separated by 0-13 nucleotides [18]. p53 binding sites may reside upstream or down stream to the coding region of target gene (ORF) or within one of its introns. p53 can also repress the transcription of some genes. This mechanism is unclear yet, but it is known that it does involve direct binding of p53 to the DNA of the repressed genes.

The human p53 contains 393 amino acids and has been divided structurally and functionally into several domains as shown in Fig. 4.2 [17]. **1.** Transcription activation domain. **2.** The sequence-specific DNA-binding domain**. 3.** The tetramerization domain.

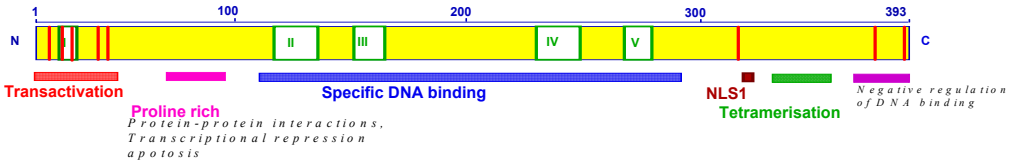

Figure 4.2. p53 has been divided structurally and functionally into 3 main domains; Transactivation domain (red), DNA binding domain (blue) and the tetramerisation domain (green).

The *transcription activation* domain of p53 is localized in the N-terminal 42 amino acids. A region within this domain (amino acids 13-29) also interacts with the human MDM2 protein [19], which regulates p53 exporting to the cytoplasm and its degradation. The *sequence-specific DNA-binding* domain is localized between amino acids 94 and 292. This domain folds into a β-sheet sandwich that forms a scaffold for a loop-sheet-helix motif and a large loop, which interacts directly with DNA. More than 90% of the missense mutations of p53 found in cancer are found in this region [6]. These mutants are defective in DNA binding, and consequently, are incapable of transactivation. The *tetramerization* domain is located in the C-terminal region, from amino acid 324 to 355 [20]. In order for p53 to be functional it has to be a tetramer in solution, and mutation in this domain may inactivate p53. The p53 protein binds to specific sequences and regulates gene expression due to its transcriptional activity. The transcriptional programs induced by p53 are heterogeneous in various cell lines or tissues, and it is likely that some of the p53 induced genes may be secondary to its primary effect as a transcription factor [21].

In summary, cells lacking functional p53 suffer from increasing of genomic instability,increasing accumulation of mutations, gene amplification or deletion phenomena, andvarious types of chromosomal aberrations. All these DNA defects can with high probability transform normal cells to malignant ones (Fig. 4.3).

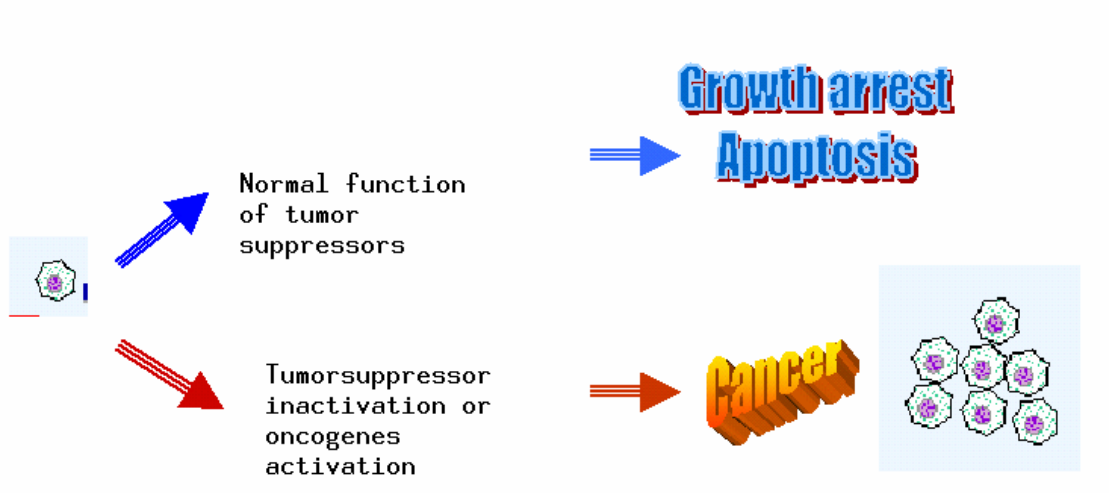

Figure 4.3. Cells lacking functional p53 suffer from increasing of genomic instability, increasing accumulation of mutations, gene amplification or deletion phenomena, and various types of chromosomal aberrations. All these DNA defects can with high probability transform normal cells to malignant ones

4.2 The experimental system

In order to analyze the transcriptional programs induced by p53, we used the temperature sensitive p53 (denoted Val135) [22], expressed in the human lung cancer cell line H1299 that lacks endogenous p53. This p53 mutant, carry a substitution from alanine to valive at position 135, behaves as a mutant at 37°C, similarly to other p53 mutant. Shifting the temperature to 32°C causes p53Val135 to undergo conformational change and to assume wild type conformation capable to induce target genes like wt p53.

Induction of growth arrest by active p53

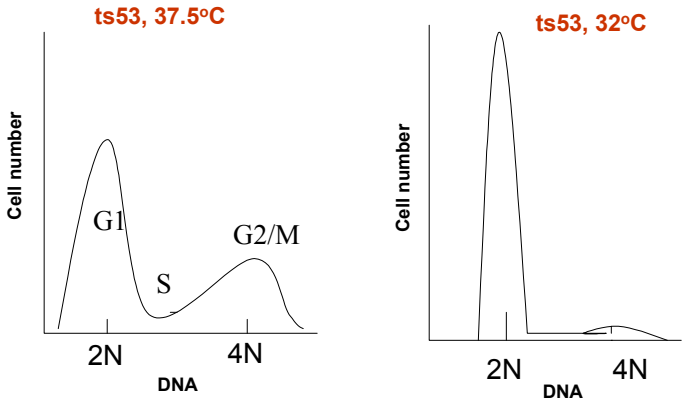

Figure 4.4. Val135p53 behaves like a p53 mutant at 37°C; no growth arrest at the G1 phase. At 32°C it behaves like wild type p53; regulates growth arrest at the G1 phase.

This conformational change does not require protein synthesis and allows for the analysis of p53-induced genes in the presence of protein synthesis inhibitor that prevents secondary effects brought about by the activated genes.

My main goal was to identify the primary targets of p53. In order to achieve this goal, I used cycloheximide (CHX), a protein synthesis inhibitor, to prevent secondary gene regulation, which is not transactivated directly by p53. Cycloheximide neutralizes the ribosome function by blocking the peptidyl transferase enzyme that is responsible for peptide bond formation during protein synthesis (see Fig. 4.5).

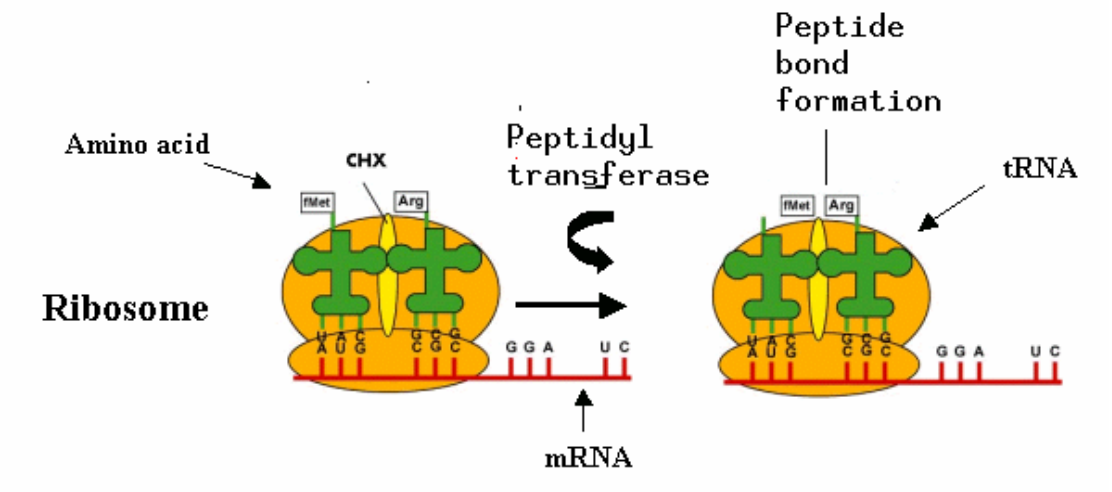

Figure 4.5. Cycloheximide neutralizes the ribosome function by blocking the peptidyl transferase enzyme that is responsible for peptide bond formation during protein synthesis.

4.3 Results and discussion

4.3.1 p53 Target genes Experiment

In this section I analyzed the profile of gene expression regulated by p53 at 32°C in the presence or absence of cycloheximide (CHX) using DNA microarrays containing ~7000 human genes and ESTs (expression sequence tags) (Genechip HuGene FL Array, Affymetrix, Santa Clara, USA). The molecular biology part of the experiments was done in collaboration with the Dept. of Molecular Cell Biology.

P53val135-based vector was incorporated and expressed in the human lung cancer cells H1299 lacking endogenous p53 and denoted val135. The control cell line used was the parental H1299 cell without the ts-p53val135, therefore p53 protein are totally absent. The H1299 cells were maintained at 37°C in RPMI medium containing 10% fetal bovine serum (FBS). The temperature was shifted to 32°C where the ts-p53 become active and total RNA was isolated, at 2,6,12,24 hrs from H1299val135 cells and for 2,12 hrs From H1299 controls cells, which lack completely p53 protein expression. The same procedure was applied to H1299val135 cells, which were exposed to cyclohexamide (10µg/µl) 30 min prior to temperature shift. The cells were harvested only up to 12 hrs (after the shifts to 32°C) at 2,4,6,9 and 12 hr time points, and total RNA was isolated.

RNA from various time points (between 2 to 24hr) was used to prepare the cRNA probe and hybridized to oligonucleotide microarrays (Genechip Hugene FL array, Affymetrix, Santa Clara, USA) which contain probes for ~7000 human genes.

4.3.2 The effect of cycloheximide on p53 target genes

To identify the primary targets, I analyzed the effect of p53 in the presence and absence of cycloheximide (CHX) at 32°C in both H1299Val135 and H1299 control cells. The inhibition

of protein synthesis by CHX presumably prevents most of the secondary gene regulation [23-24] that is not transactivated directly by p53.

It was shown that in the presence of CHX, p53 remained stable for at least 12hr and induces significantly the mRNA of *p21waf*, a major target for p53, and that protein synthesis was indeed shut down as no p21waf or hDM2 (human MDM2) proteins were detected, in contrast to their presence in the experiment without CHX [25]. It was also shown, that at 32°C p53Val135 is relocalized to the nucleus [26] and that it may therefore be protected from degradation in the presence of CHX due to lack of nuclear exclusion and the absence of hDM2 synthesis[1] [27-28].

4.3.3 The original data set

The data set is composed from gene expression values of two sets of DNA chips, 4 chips are related to cells that express p53 Val135 without CHX, and 5 chips are related to cells that express p53 in the presence of CHX. Two more chips were related to the corresponding controls.

Each chip contains probes for 7129 human genes. Each gene is represented by its expression levels taken at 9 different time points (in the two experiments combined).

The expression data can be described as a matrix *M*. Denote by *Mij* the ratio of the expression of gene *i* (where *i*=1,…,7129), measured at experiment (and time) *j* (with j=1,2,…,9), with respect to the control. To analyze the genes, we view each row of *Mij* as a vector in a 9 dimensional metric space.

4.3.4 p53 primary target genes identified in the presence of cyclohiximide

Those genes that are transcriptionally regulated by p53 under both conditions, *i.e.* with and without CHX, are likely to be primary targets for p53-mediated transcription. In order to eliminate noisy data in the analysis of the hybridization experiments, I applied a very stringent filter; I selected genes that showed more than 2.5 fold induction or repression (over their controls) at three or more time points in the presence or absence of CHX. 256 upregulated genes passed this filter for the experiment in the absence of CHX whereas 128

[1] The hDM2 is known to drive p53 to degradation.

genes passed this filter for the experiment with CHX; 38 genes were found to be common to these two groups (see the Venn diagram in Fig. 4.6). Denote by *G(38)* the group of 38 genes that have been identified by my analysis as possible primary upregulated targets of p53.

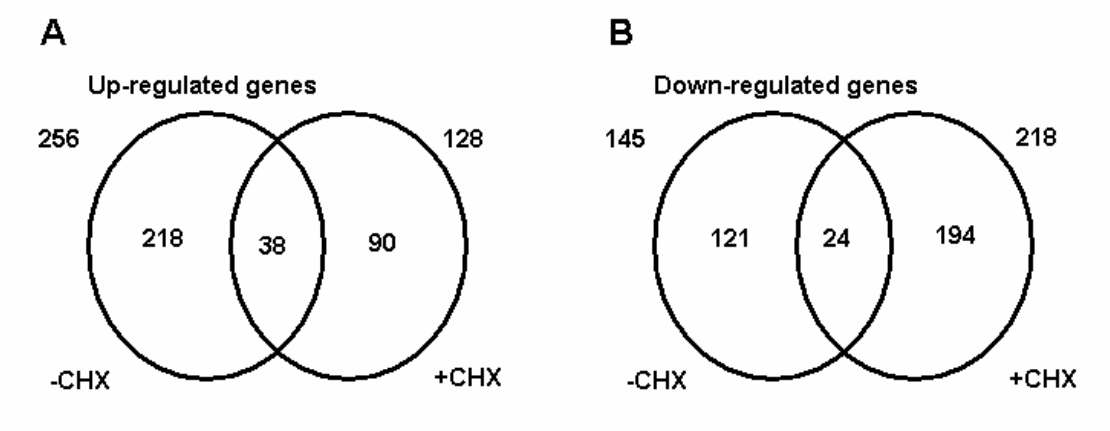

Figure 4.6. The Venn diagram represents the number of genes that were regulated by p53 in the presence and absence of CHX. (A) Only genes that showed at least 2.5 fold up or (B) down-regulation in at least 3 of the time points in each experiment *i.e.* with or without CHX, were listed in this analysis. Note that only 38 of the up-regulated and 24 of the down-regulated genes were unaffected by CHX.

I supposed that the group of genes that change their expression in the experiments with CHX (128 for up regulated and 218 for down regulated) should contain the p53 primary targets genes because of the inhibition of secondary effect of CHX. The Venn diagram (Fig. 4.6), shows, that most of the genes that were induced or repressed in the experiments with CHX (Fig. 6, 70% and 89% respectively for up and down regulated genes), have no relation to p53 target genes since they did not pass the filter for p53 regulation in the absence of CHX. I assume that the effect of CHX and temperature are the main explanation for the expression change of these genes. The genes from the experiments without CHX, that did not pass our stringent filter criteria, 218 up regulated and 121 down regulated genes, may be indirect or secondary targets of p53.

4.3.5 Are the G(38) and G(24) numbers statistical significant?

It is important to note that N=38, the number of these genes, exceeds significantly the number N_r, that would have been obtained had we applied the same filtering procedure to random data. To estimate N_r, assume that for every gene in any single experiment, the expression level can exceed 2.5 fold with probability p, and that such over-expressions are independent random events. All together, 9 measurements were taken for 7070 genes; the number of occurrences of 2.5-fold over-expression was 3364 (out of $9 \cdot 7070$). Dividing this number by all the possibilities, 63630, I estimate $p \approx 0.053$ (3364/63630). Denoting N, M as the number of experiments and successes (expression above 2.5 fold), respectively, the equation to calculate the probability that a particular gene will be over-expressed at least M times at a level above 2.5 in N experiments is:

$$P(M;N,p) = \binom{N}{M} p^M \qquad (1)$$

According to (1), the probability that a particular gene will be over-expressed at least 3 times above 2.5 in the 4 experiments without CHX is $5.7 \cdot 10^{-4}$; the same figure for the 5 experiments with CHX is $1.4 \cdot 10^{-3}$; and for passing the filter in both experiments $8 \cdot 10^{-7}$ ($5.7 \cdot 10^{-4} \cdot 1.4 \cdot 10^{-3}$). Multiplying these probabilities by the total number of genes tested, I get estimates for the numbers of genes that would have passed the filter and would have been assigned to the three groups. Next to these estimates, which represent the expected numbers of genes for random expression data, i.e. for a randomized expression matrix, I placed in parentheses the actual numbers, obtain by my analysis for the real data: with CHX: 10 (vs. 128); without CHX: 4 (vs. 256); and the number of genes expected to be in the overlap; $N_r = 5 \cdot 10^{-3}$ (vs. 38). The large disparity between the numbers obtained under the assumption of a random process and the actual measured numbers proves beyond doubt the statistical significance of my findings.

4.3.6 Identifying primary p53 target genes by cluster analysis

A major problem I must now face is that the identities of the genes of *G(38),* that were designated above as possible primary upregulated targets of p53, are determined by my stringent filtering criteria. Had I set my threshold at, say, observing 2-fold increased

expression (instead of 2.5) at two time points (instead of three), the number of genes that passed the filter would have been larger. The values of my filtering parameters (**2.5**-fold change at **3** or more times, in **both** experiments) were chosen in an arbitrary fashion; it is important to assess the extent to which relaxation of my filtering criteria will add primary p53 target genes beyond the 38 candidates that were already found. In order to reduce the dependence of our results on the precise values of the arbitrarily chosen filtering parameters, I performed cluster analysis on the data, using more relaxed filtering criteria to select the genes to be included in the analysis.

The first cluster analysis was done on the 256 genes that passed our filter for the experiment without CHX. That is, in our *filtering* we relaxed completely the restriction on the expression levels measured in the experiment with CHX (but did use the results of these experiments in the cluster analysis). Each of the 256 genes is represented by its expression levels taken at 9 different time points (in the two experiments combined). The data were normalized as follows. Denote by A_{ij} the $\log_2$ ratio of the expression of gene i (where $i=1,2,...,256$), measured at experiment (and time) j (with $j=1,2,...,9$), with respect to the control. For $j=1,...,4$ I divided A_{ij} by $\left[\sum_{j=1}^{4} A_{ij}^2\right]^{1/2}$ and for $j=5,...,9$ by $\left[\sum_{j=5}^{9} A_{ij}^2\right]^{1/2}$; the resulting 9-component vector represents gene i. The 256 genes were clustered by SPC [12-13] (see Fig. 4.7). Genes with similar expression profiles (over the time courses of both experiments) are represented by two nearby vectors and are placed in the same cluster. This cluster analysis answers directly two questions:

Do all, or a majority of the genes of *G(38)* cluster together?

What other genes cluster together with these possible primary targets?

If the answer to the first question is positive, I can identify an expression profile which is characteristic of primaries, and identify additional genes that share this profile and cluster together with the genes of *G(38)* as good candidates for being primary target genes (even though they did not pass the original stringent filtering process). Furthermore, if I find that some of the members of *G(38)* have significantly different expression kinetics than this characteristic profile, these genes should possibly be removed from the list of primaries.

Regarding the first question - the genes of *G(38)* are special in that their expression increased at least 2.5 fold at three or more time points of both experiments. Hence their representative 9-component vectors are likely to be close - but some may also be uncorrelated. For example, a gene whose expression decreases with time will have low correlation with one that increases. This situation may happen even if both genes passed the filtering criteria.

The results of my cluster analysis are summarized in the dendrogram of Fig. 4.7A. The parameter *T* on the horizontal axis controls the resolution at which the data are viewed. At *T=0* all 256 genes are in a single cluster; as *T* increases, large groups split into smaller ones. The boxes indicate clusters that contain more than 4 genes. Each box is colored according to its "purity" - the percentage of members of *G(38)* among the genes contained in the corresponding cluster. When I reorder the genes according to their position in the dendrogram, i.e. rearrange the rows of the expression data matrix according to the order imposed by the clustering process, the color-coded-matrix of Fig. 4.7B is obtained.

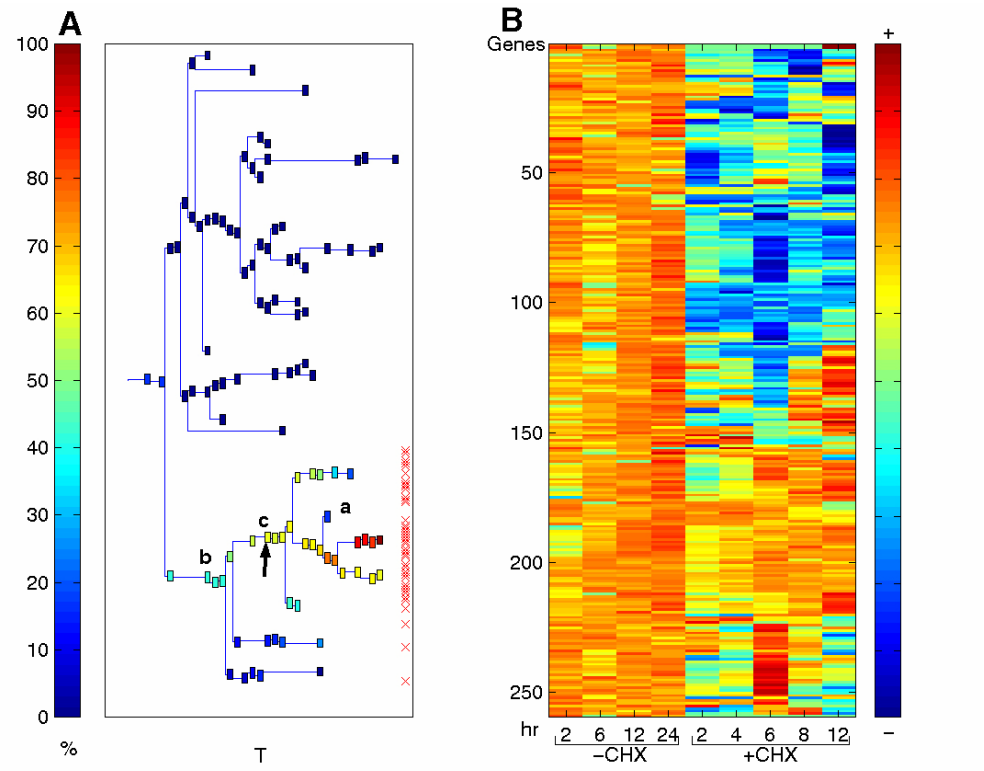

Figure 4.7. Clustering results using super-paramagnetic clustering (SPC) for the 256 genes that were up regulated at three time points or more upon activation of p53 in the experiment without CHX. (A) The dendrogram of the genes that include clusters of size 4 and larger. Each cluster is represented by a box colored according to the percent of primary target genes (38 genes, see Table 1) contained in the cluster. Red crosses at the right mark the distribution of the 38 primary target genes. (B) The normalized log ratio of the nine experiments (four without CHX and five with CHX) are plotted. The genes are ordered according to the dendrogram on the left. The color represents induction (red) or repression (blue). T, a parameter of the SPC algorithm that controls the resolution at which the cluster is found. %, Percent of primary target genes in the cluster. The cluster marked by an arrow (c) contains 87% of the 38 primary genes. The cluster marked by b contains all the 38 primary genes and the cluster marked by a contains the 9 genes that show different kinetics (Table 1 and Fig. 4)

Next, I marked the positions of the members of *G(38)* by red crosses. All 38 are in the low-level cluster denoted by *b*, which, however, contains also 58 additional genes. This cluster branches and breaks into sub-clusters, which have higher percentages of genes from *G(38)*. In order to identify a characteristic primary expression profile I want to work with a cluster, which has many members of *G(38)* (for better statistics) and also has a high percentage of them. These two requirements conflict; as I move up on the dendrogram, the clusters become purer, but also decrease in size. Hence I decided to start with the cluster *c* marked by the arrow as our working point. It contains 33 of the *G(38)* genes and, in addition, 23 genes that did not pass the original filter but have expression profiles that are similar to those of *G(38)*.

The average expression kinetics of the genes of this cluster in the two experiments is shown in Fig. 4.8A. Note the fairly similar kinetics with and without CHX, with the expression level increasing monotonously with time. All but 9 genes of *c* share these features of the expression kinetics. The 9 which differ appear on the dendrogram in the vicinity of the cluster denoted by *a*; two of these belong to *G(38)*. The average expression kinetics of these 9 genes is shown in Fig. 4.8B; it clearly differs from that of Fig. 4.8A, and their kinetics with and without CHX are different. Hence we decided to discard these 9 genes from our list of designated primaries. The average expression kinetics of the remaining 47 genes is very similar to Fig. 4.8A, with reduced scatter (error bars). The group of these 47 genes (33+23-7), denoted *G(47),* constitutes our final set of proposed primary p53 targets (Table 4.1).

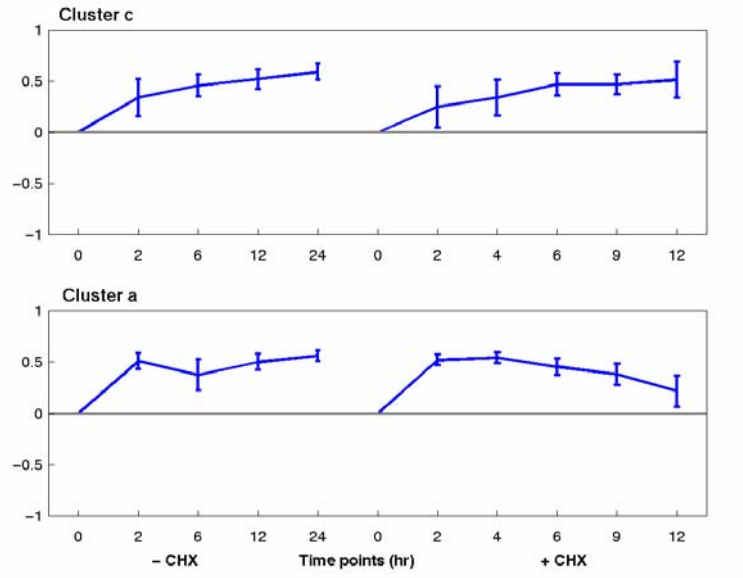

Figure 4.8. Average expression profiles of genes in clusters c and a of Fig. 4.7A. The expression profile of each gene in the cluster was normalized as described in the text.

This analysis identified a characteristic kinetic profile of primary p53 targets in either the presence or absence of CHX; the group of genes that share this profile contains 31 out of our 38 original candidates and 16 additional ones, that happened to fail our stringent filtering criteria. The various genes are listed and the groups to which they belong are properly identified in Table 4.1.

Some of the known p53 target genes such as gadd45 [32] and PCNA [31], were now included as primary targets in addition to the original 38 primary target genes, indicating that this is a sensible way to "fish" for further potential primary targets. Most of these added genes exhibited ~2 fold induction at several of the time points both in the presence or absence of CHX experiments and were previously not known to be p53 target genes.

Next, I have put the stability of our identification of primary p53 targets against changing the procedure and parameters of selection to an extremely demanding test. I performed a similar cluster analysis on a much larger set of genes, including now all those that were up-regulated at least 2-fold, at least once in the experiment without CHX. This very relaxed criterion selected 1090 genes to cluster; that are, induced 4 fold as compared to the previous clustering analysis of 256 genes; the number of genes was increased 4-fold, including now extremely noisy expression data. The results obtained when these 1090 genes were clustered are presented in Fig. 9. To my satisfaction, I found that 3/4 of our 47 proposed primary p53 targets (that were identified above), belong to two very stable gene clusters, denoted by arrows on Fig. 4.9A. The left of the two contains 24 of the genes of *G(47)* and 20 new genes, whereas the right one contains 11 from *G(47)* and 7 new associated genes. The average expression kinetics of the two clusters is shown in Fig. 4.10.

It is important to understand that adding all these extra noisy genes to the set of 256 that were analyzed before could well have resulted in a total loss of the signal that was identified above. The fact that the two clusters that are rich in previously identified primaries indeed contain 35 out of the 47 is gratifying and indicates the stability of the method used in my analysis.

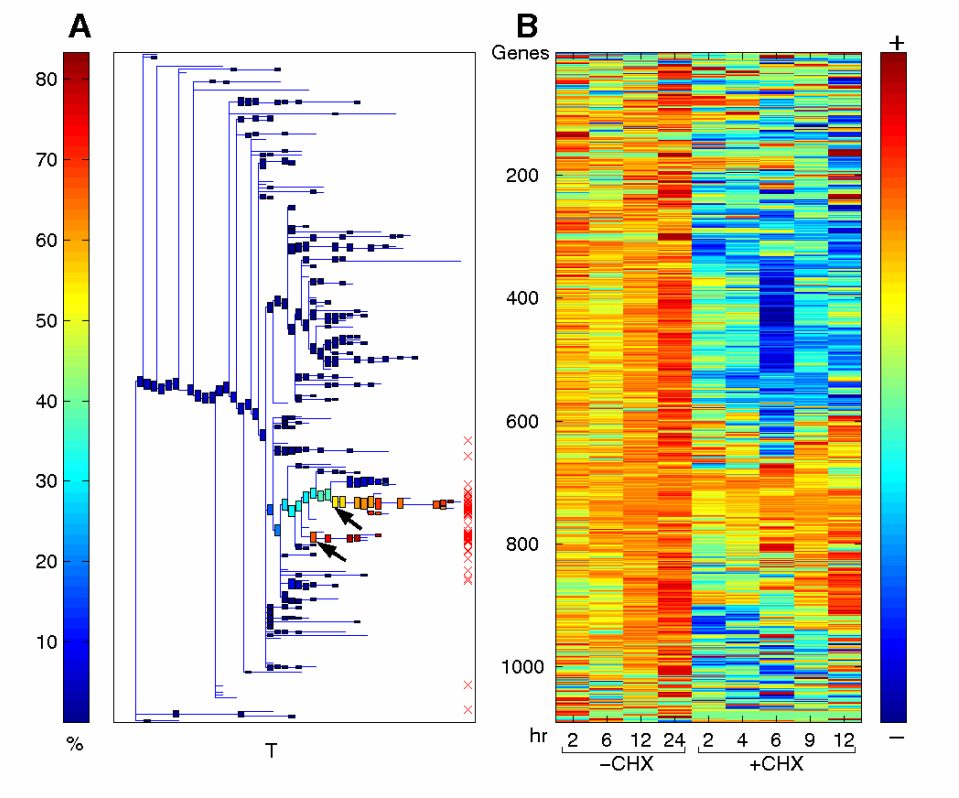

Figure 4.9. Clustering results using super paramagnetic clustering (SPC) for 1090 genes that were upregulated in the experiment without cycloheximide.
In this analysis a relaxed filtering condition was used and all the genes that were upregulated 2.5 fold at least once (1090) in the experiment without CHX were included. (A) Dendrogram of the genes including clusters of size 5 and larger. Red crosses at the right mark the distribution of the 47 primary genes. (B) Normalized log ratio of the nine experiments are plotted. The color represents induction (red) or repression (blue). Other details as in Fig. 4.3. Note the primary gene-containing cluster resolves into two distinct clusters (marked by arrows) by splitting away the non-primary gene containing clusters.

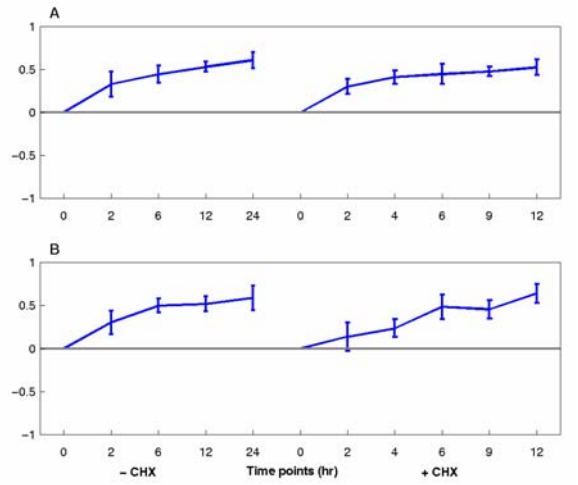

Figure 4.10. An average expression profiles of genes in clusters c and a of Fig. 4.4A. A, genes from the left arrow cluster. B, genes from the right arrow cluster (Fig. 4.4A). The expression profile of each gene was normalized as described in the text.

Finally, I discuss the choice of the clustering algorithm to be used (SPC) [12-13]. The optimal algorithm for analysis of gene expression data should have the following properties: 1) the number of clusters should be determined by the algorithm itself and not externally prescribed (as is done for SOM and K-means) [33-34]; 2) the results should show stability against noise; The method should generate a hierarchy (dendrogram) and providing a mechanism to identify in it robust, stable clusters; 4) ability to identify a dense set of points, that form a cloud of an irregular, non-spherical shape, as a cluster.

SPC, a hierarchical clustering method recently introduced by Blatt et al. (1996), is the algorithm that fits best these requirements.

Table1. Primary target genes up regulated by p53

Accession No.					Ratio of gene expression at specific time points								
					-CHX				+CHX				
	hr.				**2**	6	12	24	2	4	6	9	12
	Apoptosis	G(47)	G(38)	G(9)									
X63717	Fas/APO-1 cell surface antigen	+	+		2.2	5.6	7.9	8.3	3.2	6.8	3.3	5.7	9.8
U82987	Bcl-2 binding component 3 (bbc3)	+	+		6.2	16.5	16.8	8.4	7.6	13.4	28.5	32.5	30.5
U00115	Bcl-6	+	+		2.5	3.3	5.8	6.5	3.7	4.7	6.1	6	6.8
U16811	Bak	+	+		4	7	4.4	6.8	1	2.3	4.5	4.7	7.4
	Cell Cycle												
U09579	p21 WAF1	+	+		4.3	17.98	16.9	12.97	3.5	6.7	9.2	8.2	7.4
D90070	ATL derived PMA responsive peptide	+	+		4	4.5	2.8	4.1	1.8	3.5	3.4	4.3	6.1
M60974	GADD45	+			2.3	3.2	3.4	3.6	1.6	1.6	2.1	2.1	2
	DNA Repair/Replication												
U72649	BTG2	+	+		3	4.9	4.8	5.2	11	15.9	20.1	17.9	16.5
U18300	Damage-specific DNA binding protein	+	+		1.3	4	5.4	7.8	1.1	2.6	8.6	6.1	9.4
U90551	Histone 2A-like protein	+	+		2.6	2.6	2.5	3.5	1.7	3.1	4.5	8.8	16.7
M15796	PCNA	+			2.3	3.9	5.3	5.3	1.7	1.4	1.7	1.9	3.1
	Receptors/ECM												
X72012	Endoglin		+		21.7	15.2	7.9	6.8	3.4	1	14	6.9	1
U16306	versican	+	+		1	5.2	11.9	16.6	1	2	2.7	6.8	13.3
M21904	Heavy chain 4F2	+			7.7	6.5	8.2	6.1	0.9	1	1.6	1.3	1.4
AF010193	SMAD7			+	3.6	1.9	3.9	3.1	1.9	1.8	1.5	1.5	1.6
	Growth Factors/Inhibitors												
AB000584	TGF-Beta Superfamily Protein	+	+		3.5	41.6	67.2	50.1	11.5	16	41.4	33.6	32.4
M62402	IGFBP6	+	+		17.5	17.68	18.5	17.6	22.3	13.9	27.8	25.4	28.95
L42379	Quiescin/QSCN6	+			2.7	2.4	2.5	2.8	1.3	2.2	2.3	2.3	1.7

X97324	Adipophilin	+			5.2	6.3	7.1	10	1.5	1.9	1.7	1.9	3.8
U72263	Multiple exostoses type II protein	+			2.9	2.7	3.1	6.2	1.5	2	1.4	1.8	3.2
	Cytoskeleton/Cell Adhesion												
X13839	Vascular smooth muscle alpha-actin	+	+		6.5	14.6	44.3	98.3	6.4	9	18.9	18.4	26.9
Z49989	Smoothelin	+	+		3.7	4.3	3.9	3.5	1.7	2.2	4.6	4.1	3.4
X05608	Neurofilament subunit NF-L	+	+		3	7.7	16.6	32	5	8.8	9.6	16.8	32.2
D82345	NB Thymosin beta	+	+		2.4	3.4	4.2	5.4	1.6	2.5	3.8	2.8	3.9
X93510	LIM domain protein		+		4.5	5.5	4.3	5.6	4.3	2.4	9.3	3.8	1
	Metabolism												
U24389	Lysyl oxidase-like protein	+	+		16.7	17.5	23	22.9	5.5	4.2	9.7	8.3	6.9
L41668	UDP-Galactose 4 epimerase (GALE)	+	+		4.2	3.6	3.2	4.4	3.2	3.7	5.1	3.4	2.8
Y12556	cAMP activated Protein Kinase B	+	+		4	1	19.7	22.2	1	3.1	3.1	7.1	37.5
U05572	Lysosomal Mannosidase alpha B		+		9.2	2	7	15.9	6.5	5.6	13.9	13	7.2
Y09616	Carboxylesterase (liver)	+	+		2.3	4.1	5.9	9.1	1.9	2.5	3	3.4	3.6
U78735	ABC3			+	4.5	3	2.8	3.6	1.9	2	1.9	1.4	1.1
M20902	Apolipoprotein C-I (VLDL)	+			3.1	3.9	2.9	4.1	2.1	2.9	2.3	2.2	1.7
U20325	CART	+			2.8	3.9	3	3.3	1.3	1.7	2.1	1.3	1.9
M12625	Lecithin-cholesterol acyltransferase	+			1.4	2.9	3	3	0.8	1.4	3.6	2.6	2.4
D87292	Rhodanese			+	7.7	4.9	5.4	10.5	1.6	1.6	1.7	1.3	1.2
	Neuronal Growth												
U35139	NECDIN related protein		+		3.2	1	3	15.7	5.8	4.9	2.7	4.8	5.5
M96740	NSCL-2 gene	+	+		1	4.2	10.1	12.6	2.6	2	4.8	7	11
U60062	FEZ1-T gene	+	+		1	2.7	2.7	4	1.6	2.3	2.6	5.6	8.7
U72661	Ninjurin 1	+	+		1.4	3.7	7.6	10.4	1.9	3.1	6	6.3	7.8
U48437	Amyloid precursor like protein	+	+		1.9	2.7	3.7	4.7	1	1.5	3.6	2.7	3.8
	Signal Transduction												
J00277	c-Ha-ras 1	+	+		2.3	3.1	4.7	4.8	1.2	1.5	4.1	5.6	5.7
X77777	Intestinal VIPR related protein		+		14.7	7.5	11.9	9.8	1	2.7	19.9	4.7	1
X62535	Diacylglycerol Kinase (alpha)	+	+		4.4	12	11.2	38.8	1.2	3.4	6	7.1	10.1
U56998	Putative ser/thr protein kinase	+	+		6.3	3.6	1	12.5	2.1	2.2	4.8	6.4	4.5
L08835	DM Kinase	+			2.5	3.4	3.3	3.7	0.9	1.2	1.8	1.5	1.5
L42176	DRAL/FHL2			+	5.4	3.8	3.8	6.5	2	2	1.8	1.5	1.7
	Transcription												
L19871	Activating transcription factor 3	+	+		1.4	4.4	5.9	5.2	4.2	7.2	18.9	8.8	9.9
U41315	ZNF127-Xp		+	+	4.4	5.1	9.2	9.5	6.5	6.4	3.9	7.2	3.1
AD000684	LISCH7	+	+		5.2	6.1	6.1	7.4	2.7	1	2.1	3.1	3.8
M29580	Zinc finger protein 7	+			1.2	2.6	3.6	4.7	1.2	1.2	1.2	1.3	1.2
U90913	Tip-1	+			3.4	2.7	3.1	3.7	1.5	2.1	2.3	1.9	2.3
HG3494	Nuclear factor NF-116			+	3.3	2.4	2.9	2.7	2.9	2.8	2.4	2.3	1.1

	Other												
U10099	POM-ZP3	+	+		3.2	4.2	4.7	4.8	1.8	2	4	3.5	3.3
D87434	KIAAA0247	+	+		1	4.5	4.7	6.6	3.5	7.5	10.3	15.4	18.9
U33147	Mammaglobin 1		+	+	3.5	1	4.5	4.1	4.5	4.2	2.9	4.6	1
J05016	Disulfide isomerase related protein	+			2.6	3.8	6	8.7	2.1	1.4	1.8	1.8	2.3
U81556	OS4			+	2.7	2.6	2.4	3	1.6	2.2	1.8	1.4	1.3
S58544	Infertility-related sperm protein	+			3	3.9	7.6	3.1	1.6	2	1.7	3	4.3
D63481	KIAA0147	+			1.8	2.6	3.1	2.6	1.2	1.2	1.5	2.1	1.7
Z35093	SURF-1			+	2.2	2.6	2.6	3.4	1.9	2	2	1.6	1.4
U94747	WD repeat protein HAN11	+			2.3	3.1	3.4	3.4	0.8	1.3	3	1.4	3.1

Chapter 5

Breast Cancer

5.1 Introduction

Breast cancer is a major cause of death among women in the age group of 33-55 years. Despite important advances in therapy, still more than half of the patients suffer from relapses. Therefore, further molecular characterization is needed to improve diagnostic and therapeutic strategies.

5.1.1 Anatomy of the breast

The female breast is composed of 15-20 sections, called lobes, with each lobe ending in many smaller lobules. These lobules further end in dozens of tiny bulbs, which produce milk during lactation. The lobes are shown in Figure 5.A as clusters of smaller yellow circles, which represent the smaller lobules. The three components, lobes, lobules, and bulbs, are all linked together by thin tubes called ducts, which are shown as small brown tubes running throughout the breast. All the small ducts eventually come together to form larger ducts, which empty to the outside through the nipple. The bulbs are too small to be seen on this diagram.

5.1.2 Types of breast cancer

Most breast cancers form initially in the ducts. When the disease is confined to the ducts it is called intraductal cancer. These types of cancer do not have the potential of spreading and are curable. When the cancer cells break out of the duct, the cancer becomes invasive, and is called Invasive Breast Cancer (IBC). At this stage the cancer cells behave abnormally and grow haphazardly, and are capable to invade into the normal breast tissue in many areas.

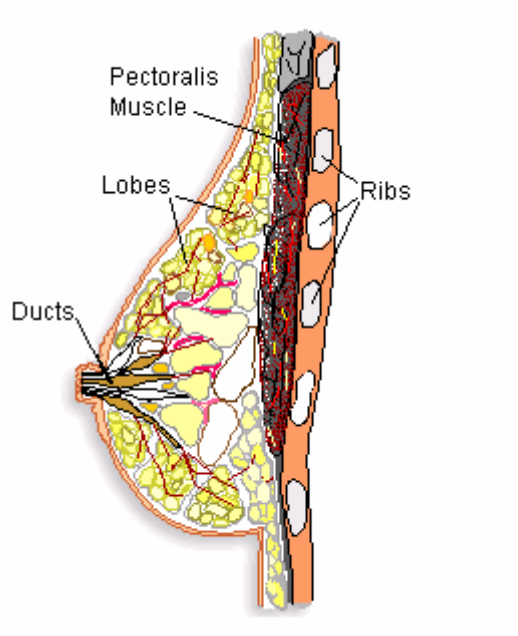

Figure 5.A. The female breast is composed of 15-20 sections, called lobes, with each lobe ending in many smaller lobules. The smaller yellow circles represent the smaller lobules. The three components, lobes, lobules, and bulbs, are all linked together by thin tubes called ducts (shown as small brown tubes). All the small ducts eventually come together to form larger ducts, which empty to the outside through the nipple (taken from the web site: http://breastdoctor.com/breast/).

A non-invasive ductal cancer will stay in the duct and not invade the surrounding breast tissue. Invasive ductal cancer affects the ducts and lobules of the breast and has the potential to spread widely. This type of cancer is considered to be the less life-threatening because it tends to stay in the same breast in which it originally occurs, but it is still a very serious breast cancer. In the case of invasive **lobular cancer,** single cells are invading into the normal breast tissues. This type of breast cancer is considered to be the more dangerous and aggressive because it has a propensity for recurring in the opposite breast. Most clinical studies have shown that seventy to eighty percent of patients have survived five years after treatment, but this survival rate is less than what is expected for the vast majority of breast cancers. At the invasive stage, the cancer cells may gain access to the lymphatic vessels. These tiny vessels carry a pale fluid that passes into bigger and bigger lymphatic channels, through lymph nodes, and eventually into the blood stream. All lymphatic channels lead to lymph nodes, small rounded masses of tissue, which filter the lymphatic fluid to keep foreign objects such as bacteria and cancer cells from gaining access to the rest of the body. The liquid portion of

lymph fluid resembles blood plasma (the clear yellow part) and contains white blood cells but no red blood cells. The presence of cancer cells in lymph nodes is an indication that cancer has spread. When the cancer cells invade further from the lymph nodes to the blood vessels, they may spread all over the body (metastasis) and the breast cancer become much more serious. This stage may recur after the primary (initial) treatment and is the type that causes the death of patients.

5.1.3 From pre-existing benign lesions to Invasive Breast Cancer (IBC)

The majority of IBCs are thought to develop over long periods of time from certain pre-existing benign (non-cancerous, or non-malignant) lesions. There are many types of benign lesions in the human breast and only a few appear to have significant premalignant potential. The best characterized premalignant lesions recognized today are referred to as "atypical ductal hyperplasia" (increased cell production in normal tissue) (ADH), "atypical lobular hyperplasia" (ALH), "ductal carcinoma *in situ"* (DCIS), and "lobular carcinoma *in situ"* (LCIS). All these lesions possess some malignant properties such as a relative loss of growth control, but they lack the ability to invade and metastases and, in this sense, are premalignant [35]. The structure of the normal epithelium in the Terminal Duct Lobular Units (TDLUs) varies considerably as a function of hormonal status (e.g. menstruation, pregnancy, etc.). The TDLUs group into four histological categories (type I through type IV) on a continuum of differentiation towards lactation. Type I TDLUs, the least differentiated, have relatively high proliferation rates and are more common in cancerous breasts, suggesting that this stage is more susceptible to growth alternations with premalignant potential. Most of the premalignant lesions (e.g. ADH, ALH, DCIS, LCIS, etc.) seem to develop to become malignant from this stage [35]. The overall growth of premalignant breast lesions can be viewed simplistically as a balance between cell proliferation and cell death. On average, the cells in all types of premalignant lesions proliferate faster than normal cells in TDLUs, contributing to their positive growth imbalance [35]. Much less is known about cell death. One preliminary study reported significantly lower rates of apoptosis in ADH compared with TDLUs in the same breast. However, a few studies have reported rates of apoptosis in DCIS that are up to 10-fold higher than typically seen in normal cells [36-37], suggesting that the relationship between

cell proliferation and death may not always be accurately portrayed by the static methods used to measure these dynamic processes.

5.1.4 Hormones, oncogene and tumor suppressor genes in premalignant breast cancer

Estrogen receptor. Estrogen, mediated by the estrogen receptor (ER), plays a central role in regulating the growth and the differentiation of normal breast epithelium [35]. It stimulates cell proliferation and regulates the expression of other genes including the progresterone receptor.

Prolonged estrogen exposure is an important risk factor for developing IBC, perhaps by allowing random genetic alternations to accumulate in normal cells stimulated to proliferate, which may also be true for cells in premalignant lesions. The very high levels of ER observed in nearly all-premalignant lesion cells might be duo to their high efficiency to respond to any level of estrogen [35]. A somatic mutated ER showed much higher transcriptional activity and proliferation than wild type. The mutated ER also showed increased binding to the co-activator TIF-2 (a nuclear steroid receptor coactivator, family member of SCR-1), which may partially explain its increased functional responsiveness to estrogen [38]. However, there may be other alternations of ER resulting in increased growth. For example, proliferation in TDLUs occurs predominantly in ER-negative epithelium, whereas the majority of dividing cells in premalgnant lesions are ER positive. These data are consistent with the hypothesis that cells in normal human breast epithelium are hierarchical in organization and support a model in which proliferation of ER-negative cells is controlled by paracrine factors released from ER-positive cells under the influence of estradiol. This organization may be disrupted in some tumors [39].

The tumor suppressor – p53. The tumor suppressor p53 also appears to play an important role in the evolution of premalignant breast disease. This tumor suppressor gene is mutated in about 30% of IBCs, which is associated with generally aggressive biological features and poor clinical outcome [48-49]. Most are missense point mutations resulting in an inactivated but stabilized protein that accumulates to very high levels in the cell nucleus. Mutations of

p53 may contribute to the development and progression of premalignant breast disease by several mechanisms, including interference with DNA repair through loss of an important G1 cell-cycle checkpoint, leading to replication of damaged DNA template and genetic instability, and also perhaps by clonal expansion through inhibition of programmed cell death [17].

The ErbB2 oncogene. ErbB2 is a receptor tyrosine kinase from the EGFR receptor family that is amplified and/or overexpressed in 20-30% of IBCs [42]. These abnormalities are associated with increased proliferation, poor clinical outcome, and altered responsiveness to various types of adjuvant therapies. ErbB2 may also promote cell motility, which could contribute to the ability of tumor cells with overexpressed ErbB2 to invade and metastasize. Just how alteration of ErbB2 leads to the development and progression of premalignant breast disease is not entirely clear, although both the increased proliferation and motility of cells associated with over expression may contribute. Whatever the mechanism, the absence of over expression in normal TDLUs and ADH, compared with relatively high rate in DCIS [35], suggests that alterations of ErbB2 are an important event in early malignant transformation. One of the new drugs against breast cancer is the antibody to ErbB2 (called herceptin) which block the function of ErbB2 and inhibit to some extent the growth of the tumor.

5.1.5 Hereditary breast cancer

Some 5-10% of cases are thought to be inherited. The hereditary breast cancer includes genetic alternations of various susceptibility genes, particularly *BRCA1* and *BRCA2*. Breast tumors of patients with germ-line mutations in the *BRCA1* and *BRCA2* genes have more genetic defects than sporadic breast tumors. Accumulation of sporadic genetic changes during tumor progression follows a specific and more aggressive pathway of chromosome damage in these individual [40]. The protein product of the *BRCA1* gene is implicated in the cellular response to DNA damage, with postulated roles in homologous recombination and in transcriptional regulation. *BRCA1* and p53 may coordinately regulate gene expression in their role as tumors suppressors. Two categories of genes are significantly altered in *BRCA1* transected cells: cell cycle control genes and DNA damage response genes. Progress in the past 5 years has offered the possibility of a molecular test for genetic screening for inherited mutations of cancer predisposing genes.

5.1.6 Gene expression analysis in human breast cancer

The study of gene expression in primary human beast tumors, as in most solid tumors, is complicated for two major reasons. First, breast tumors consist of many different cell types, including not just carcinoma cells, but also additional epithelial cell types, stromal cells, adipose cells, endothelial cells, and infiltrating lymphocytes. Second, breast carcinoma (BC) cell themselves are morphologically and genetically diverse [41]. These features have made the study and classifications of human breast tumors difficult. Recently, novel array hybridization techniques based on cDNA or oligonucleoties have enabled the parallel expression profiling of several thousand genes, providing a powerful tool for characterizing complex cellular transcription activities. At present, one major aim is to use DNA arrays as a tool to understand and classify tumors into categories based on shared gene expression patterns. Perou et al. (Stanford University, USA) showed that primary breast tumors can be classified, based on at least two different gene-expression parameters: expression levels of the genes in the proliferation-associated cluster and the interferon (IFN) regulated signaling pathway cluster [41]. They found a subclass of BC tumors that differ in the expression levels of STAT1 and STAT3, resulting in the induction of a known set of IFN-regulated genes. Latter studies [42-43] done by the same group characterized variation in gene expression patterns in breast tumors, using cDNA microarrays containing 8,102 human genes. They showed that human breast tumors could be classified to several subgroups. In the first study [42] they identified four subgroups of breast tumors that might be related to different molecular features of mammary epithelial biology; ER+/luminal-like, basal-like, ErbB2-positive and normal breast. A latter study [43] showed that the previously characterized luminal epithelial/estrogen receptor-positive group (previously denoted by ER+/luminal) could be divided into at least two subgroups, each with a distinctive expression profile. Survival analyses done in this study showed that estrogen receptor-negative classes (basal-like, ErbB2-like and normal-cell-like tumors) were all associated with poor outcome [43]. Additionally, the two estrogen receptor-positive groups show significant difference in their outcome.

Ahr A. et al. (Goethe University, Frankfurt, Germany) tested the gene expression profile of 82 specimens using cDNA arrays [44]. The major aim of this study was to identify differential expressed genes in breast cancer, which can subsequently be employed as markers for a

molecular characterization of tumor samples. Using a hierarchical clustering method, they identified four main subgroups of samples, and checked the correlation of these groups with classical clinicopathological parameters (e.g. histological subtype, grade, etc.). While no correlation was detectable between cluster data and tumor size, grade or histological subtype, one subgroup was enriched of node-positive tumors (88%). Interestingly, in this subgroup they also detected an accumulation of samples from patients who had already developed distant metastases at the time of diagnosis.

Numerous studies have correlated genetic alterations with clinical outcome, including a strong correlation between the amplification of the ErbB2 receptor gene (Her-2) and poor clinical outcome [45-46]. Nevertheless, such correlations are few and often do not adequately define tumor subtypes. West M. et al. (Duke university, USA) [47] developed a statistical method to identify gene subsets that have the capacity to discriminate breast tumors on the basis of estrogen receptor status and also on the categorized lymph node status. These groups of genes include some that function in the ER pathway, including the *ER* gene itself as well as a number of known targets for ER. Several others contribute to the discrimination inversely with ER+ status; some of these encode proteins known to have inverse relationship with ER function. A major practical interest and potential clinical value of such statistical analyses lies in the ability to provide the ER status of the tumor on the basis of gene expression profile. This may have implication on treatment and prognosis.

5.1.7 Resistance to doxorubicin in breast cancer

Chemoresistance is the main obstacle to successful therapy in cancer patients. The merging understanding of the key role of apoptosis to the effects of chemotherapy has led to a focus on defects in the apoptotic machinery as a cause of chemoresistance. P. E. Lonning et al. [48] and Geisler S. et al. [49] confirm previously studies, that TP53 mutations affecting certain domains of the p53 protein are associated with primary resistance to doxorubicin therapy in breast cancer patients. Similarly, expression of c-erbB2, a high histological grade, and lack of expression of bcl-2 all predicted chemoresistance. TP53 mutations were associated with expression of c-erbB2, high histological grade and bcl-2 negativity. Additionally, TP53 mutations, a high histological grade, and lack of expression of bcl-2 (but not TP53 LOH) all predicted high relapse. They show that certain TP53 mutations predict resistance to doxorubicin in breast cancer patients. Their findings are consistent with the hypothesis that

other defects may act in concern with loss of p53 function, causing resistance to doxorubicin in breast cancer [48]. This indicates that doxorubicin causes DNA damage, which activates p53 that activates proapoptotic genes. In the absence of p53 function doxorubicin looses its effect.

Tumors are currently diagnosed by histology and immunohistochemistry based on their morphology and protein expression, respectively. However, poorly differentiated cancers can be difficult to diagnose by routine histophatology. In addition, the histological appearance of a tumor cannot reveal the underling genetic aberrations or biological processes that contribute to the malignant process. Kahan J. et al. [50] developed a method of diagnostic classification of cancers from their gene-expression signatures using artificial neural networks (ANNs). They calibrated the ANNs using the small, round blue-cell tumors (SRBCTs) as a model. These cancers belong to four distinct categories and often present diagnostic dilemmas in clinical practice. The ANNs correctly classified all samples and identified the genes most relevant to the classification, and produced a list of genes ranked by their significance to the classification. When they tested the ANN models calibrated using the top 96 genes on 25 blind samples, they were able to correctly classify all 20 samples of SRBCTs and reject the 5 non-SRBCTs. This supports the potential use of these methods as an adjunct to routine histological diagnosis.

5.1.8 clustering analysis of gene expression data of human breast tumors

My aim is to re-analyse the data that first appeared in the paper “Molecular portraits of human breast cancer” (by Perou et al) [42]. I used the data that was made available on the website http://genome-www.stanford.edu/breast_cancer/molecularportraits/download.shtml

I posed the following questions:
1.Do our methods of analysis reproduce the results obtained by Perou et al?
2 .Can we make observations that seem to be of interest and were not reported by Perou et al?

There are several differences between our methods of analysis and those of Perou et al. These can be summarized as follows, ordered according to increasing importance and complexity.

a. Normalization: We normalize the samples differently. Whereas Perou et al center and normalize the genes as well as the samples, we do NOT center and normalize the samples.

b. The clustering method: Perou et al use Average Linkage (AVL), the method introduced to the gene expression literature by Eisen et al. We use the SuperParamagnetic Clustering algorithm (SPC) [12-13]. Using SPC is essential for our main advance - that of using the Coupled Two Way Clustering method (CTWC). The important feature of SPC is its ability to assign a stability index to each cluster in any dendrogram that we generate. Stable clusters are more significant statistically and less likely to be due to random fluctuations and noise in the data .

c. Intrinsic set vs CTWC: Perou et al noticed that if all 1753 genes (that passed their initial filtering) are used to analyze and classify the samples, important information is wiped out. The same observation was made in general by us in our PNAS paper [14]; some potentially important and meaningful partitions can be seen only if one "listens" to small subsets of genes, one subset at a time. The large majority of the genes usually do not contribute to differentiating a particular process of interest, and, even worse - they produce random noise that masks the effect of the "important players".

Perou et al proposed a way to overcome this problem by using criteria that are based on their biological insight, knowledge and bias, which leads them to prune down the genes to an "intrinsic set" of 496. When only the expression levels of these genes are used, Perou et al find that the tumors break into several classes of interest. As opposed to this approach, which is knowledge based, introduces biases into the analysis and can be used only when matched samples (e.g. BEFORE and AFTER chemotherapy) are available, CTWC is an automated method [14], which does not necessitate any biological wisdom to prune the genes

5.2 Summary of the results of Perou et al

The main goal of this paper was to develop a system for classifying tumors on the basis of their gene expression patterns. This paper characterizes gene expression profiles of 65 tumors and 19 cell lines, using cDNA microarrays, representing 8,102 human genes. In this work twenty of the 65 tumors were sampled twice; 18 from patients who were treated with doxorubicin (chemotherapy) for an average of 16 weeks, with surgical biopsy done before and

after the treatment, and two more tumors were paired with a lymph node metastasis from the same patient. The 25 remaining specimens included three normal tissues and 22 tumors. The original expression table (matrix) included 8,102 rows, each corresponding to a gene, and 84 columns, each corresponding to a sample. In order to focus on the most interesting genes, Perou et al selected the subset of genes whose expression varied by at least 4-fold from the median of the samples, in at least three of the samples tested. This filtering process left 1753 genes, each of which is represented by 84 expression values. In the final expression matrix Perou et al split the data matrix into two submatrices; one of tissues and one of cell lines. The two submatrices were, separately, median polished (the rows and columns were iteratively adjusted to have median 0) before being rejoined into a single matrix.

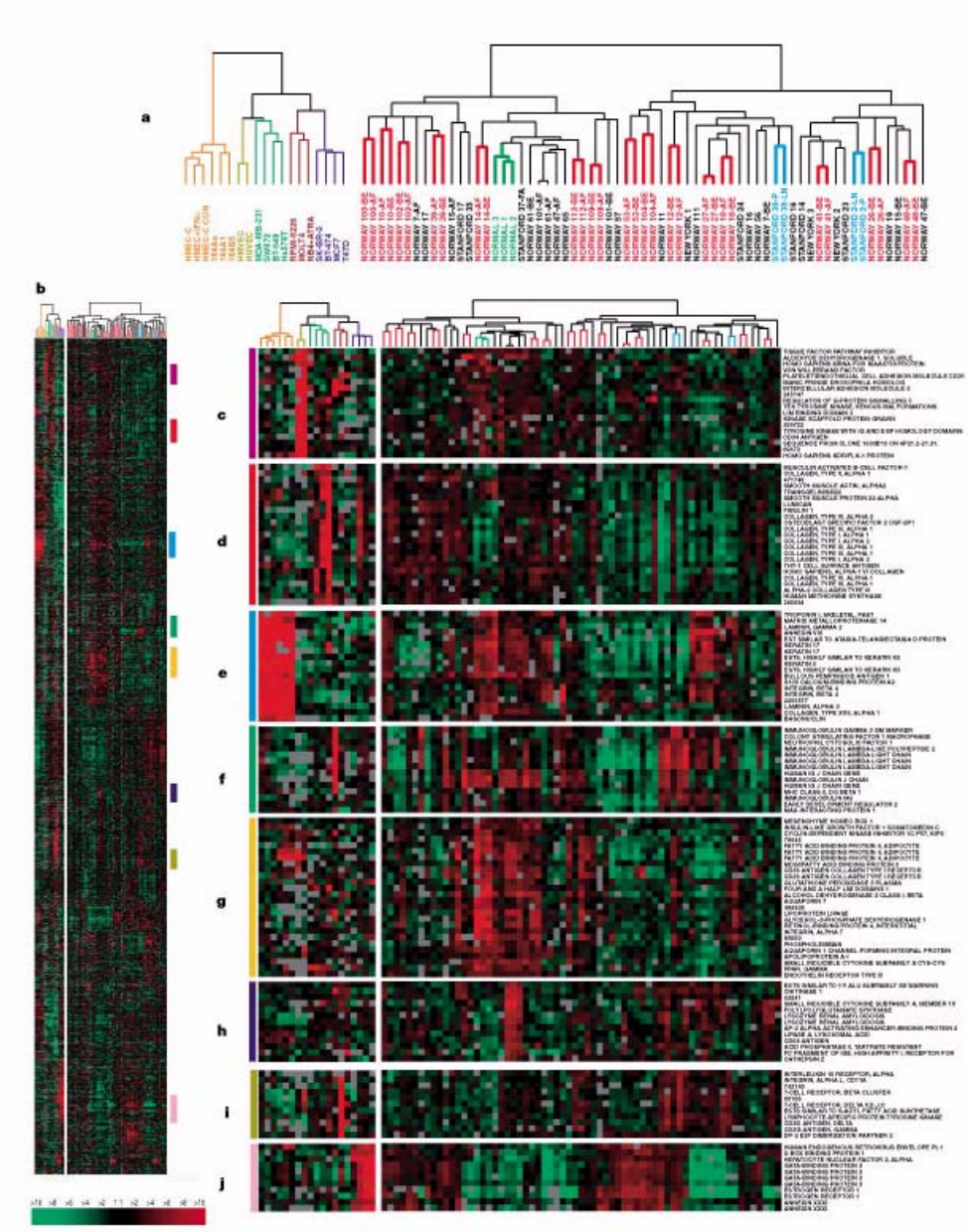

Figure 5.1. Cluster analysis of the set of 1753 genes, based on all 84 samples (Fig.1 of Perou et al).

In the first part of this work, Average Linkage (AVL), a hierarchical clustering method was used to cluster the 1753 genes on the basis of pairwise similarities of their expression profiles over all the 84 samples. The same clustering method was used to cluster the samples, the 19 cell lines and 65 tissues separately, on the basis of the similarities of their expression profiles over all the 1753 genes. Fig. 5.1 presents the results of this analysis (reproduced from Perou et al) [42].

The expression matrix was two-way clustered; clustering the genes was based on the 84 samples, and clustering the 84 samples was based on the 1753 genes. The expression matrix was reordered according to the two-way clustering, and is shown in Fig. 5.1. Normalizing and clustering the 19 cell lines and the 65 tumors, separately, using the AVL method, generated the dendograms presented in Fig. 5.1a. Figs. 5.1c-j presents 8 clusters of genes, which Perou et al found to be the most interesting. Another cluster, named "proliferation", is also mentioned and reported in the supplementary information (http://genome-www.stanford.edu/breast_cancer/molecularportraits/figures.shtml).

Referring to the results shown in Fig. 5.1, Perou et al mentioned three striking features;

The tissues show great variation in their patterns of gene expression. Evidence for this statement is shown in the sample to sample distance matrix (see Fig. 5.2), with the samples ordered according to the dendrogram shown on the right hand side of Fig. 5.1a. One sees that most of the tumors are at large distances from one another (red), except for several small groups (represented by yellow and blue and identified by arrows), whose relative pairwise distances are closer to one another than to the other samples.

Different sets of genes show unique patterns of variation over the samples. For example, consider the set of genes represented in Fig. 5.1j; one sees that this set divides the tissues to groups that exhibit either low or high expression levels (represented by the green and red colors, respectively). These patterns highlight various relationships among groups of genes, among the tumors, and various connections between specific genes and specific tumors.

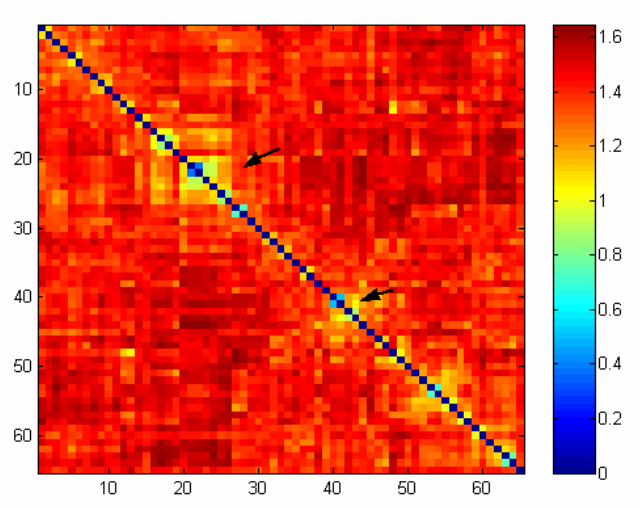

Figure 5.2. This matrix represents the distances between the 65 tissues, based on the expression profiles of the 1753 genes after normalize the samples. This distance matrix corresponds to the dendrogram on the right hand side of Fig. 5.1a. Red color represents large distance and blue color represents small distance between the samples. The two arrows point to two groups of tissues, which are closer to each other than to other tissues. This distance matrix was generated after clustering the samples and reordering them accordingly. Looking at the dendrogram in Fig. 5.1a (right hand side) one sees a clear partition to two clusters of samples. From the features seen from the distance matrix it can be concluded that this partition is not as significant as implied by the dendogram.

The findings in Fig. 5.1 demonstrated that the 1753 genes were not an optimal subset to classify the tumors on the basis of their gene expression patterns. Therefore Perou et al turned to selecting a subset of 496 “intrinsic” genes that consists of genes with significantly greater variation in expression *between different* tumors than *between paired* samples from the same tumor. The analysis, done in the second part of the paper, was based on these 496 genes, and its results are presented in Fig. 5.3 (which corresponds to Fig. 5.3 of Perou et al).

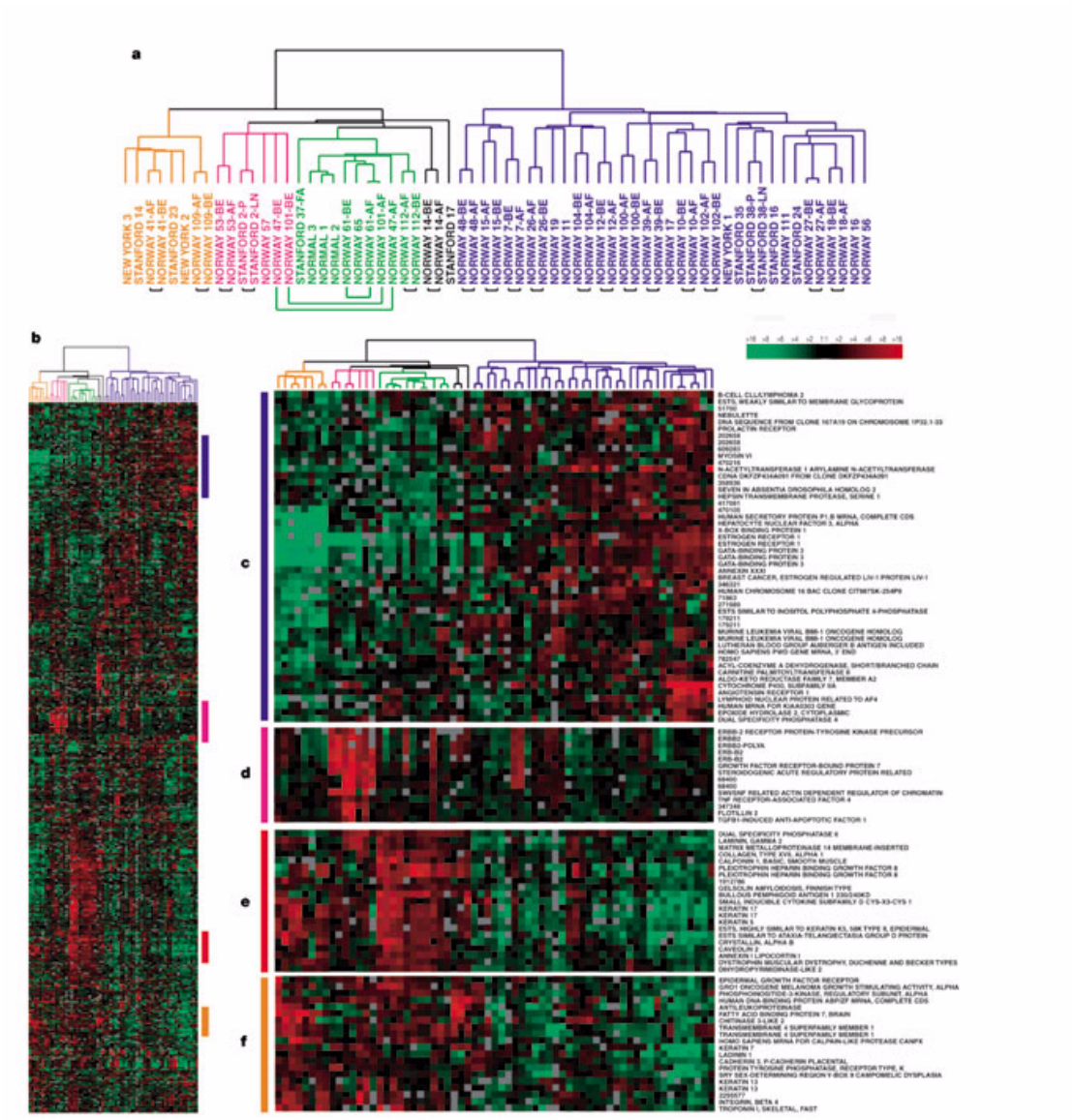

Figure 5.3. Clustering the samples using the 'intrinsic' gene subset, and of the 496 genes, using all 65 tumors. (Fig.3 of Perou et al)

At this stage of the analysis Perou et al used a data matrix of the expression level of these 496 genes, in each of the 65 tumors.

This matrix was two-way clustered, using the AVL method, like in the first part. Here again, the tumors were normalized before clustering. The expression matrix, reordered according to the clustering is presented in Fig. 5.3b.

Looking at Fig. 5.3, one sees that Peoru et al classify the 65 tumors to four groups, identified in the dendrogram of Fig. 5.3a as ER+/luminal-like (blue), basal-like (yellow), Erb-B2+ (red) and normal breast-like (green).

5.3 My two way cluster analysis and comparison

In my work I analyze the same data set that was used by Perou et al to obtain Fig. 3.1. The 3 main differences between my method and Perou et al are:

Normalization. I used the same normalization method for the genes but I did not normalize the samples before clustering them.

Clustering method. I used the Super Paramagnetic Clustering algorithm (SPC) [12-13]. Full details about this method are given in chapter 2.

Selecting sets of genes. The main difference is the way of selecting the subsets of genes to be used for clustering the tumors. I used the Coupled Two way Clustering algorithm (CTWC)[1] [14]. As opposed to selecting the 496 "intrinsic" genes, CTWC selects several subsets of genes in an unsupervised way, without using any preliminary information that is not contained in the expression matrix (such as preferring genes that show larger differences between unpaired samples).

The comparison revealed two major points; first, that the main findings shown in Fig. 5.3 can be found directly (and improved upon), starting from the entire set of 1753 genes and using the CTWC, without filtering the genes further (to the "intrinsic set"). Second, I find new tumor classifications that were not mentioned by the Perou et al.

5.4 Method

5.4.1 Clustering all Genes *versus* all Samples – G1 (S)

Denote the entire set of 1753 genes by G1, and by S the 84 samples. G1 (S) denotes the clustering operation of G1, based on S. For G1 (S) the comparison is directly between the AVL and SPC clustering methods, since we normalize the data in the same way. Denote by

[1] Short Explanation is mentioned below in the CTWC section, and more details are reported in chapter 3

E_{ij} the relative expression of gene *i* in sample j. The data consist of 1753 points in a 84-dimensional space, normalized in the standard way:

$$G_{ij} = \frac{E_{ij} - \langle E_i \rangle}{\sigma_i}, \qquad \langle E_i \rangle = \frac{1}{84}\sum_{j=1}^{84} E_{ij}, \qquad \sigma_i^2 = \frac{1}{84}\sum_{j=1}^{84} E_{ij}^2 - \langle E_i \rangle^2. \qquad (1)$$

With this normalization, the squared Euclidean distance between genes i and j is a linear function of their Pearson correlation coefficient.

The comparison between AVL and SPC was based on the gene clusters that are shown in Fig. 5.1 (c to j and the proliferation cluster, reported in the supplementary information of Perou et al). For each cluster in Fig. 5.1 we identified the homologous cluster given by SPC, as the one with the highest Purity (P) and Efficiency (E) (see Table 1 in the Results section). Denoting by X a cluster of Fig.1 and by Y a cluster of SPC,

$$E = \frac{|X \cap Y|}{|X|}, \qquad P = \frac{|X \cap Y|}{|Y|}. \qquad (2)$$

Except for one case, all the gene clusters present in Fig. 5.1c-j were identified also by SPC with relatively high purity and efficiency (more details are presented below in the Results section).

Therefore we can hypothesize that for this data there are not significant quantitative differences between the results obtained by the two methods.

5.4.2 Clustering all samples *versus* all Genes – S(G1)

Clustering the 84 samples (cell lines and tumors, separately) based on the 1753 genes was the next step of the analysis (Fig. 5.1a). In this step, in addition to using different clustering methods, we also used different normalizations. Before clustering, Perou et al normalized the samples in the same way as was done for the genes. We hold the opinion that the samples should not be normalized this way before clustering[2] and try, in the following, to explain why.

5.4.2.1 The normalization dilemma

In order to understand our reasoning, one should focus on Fig. 5.4.

[2] Initially each sample was normalized (scaled) once, over all genes.

The 4 matrices A, B, D and E represent the expression levels of 12 highly correlated genes (that form a single cluster) ordered along the horizontal axis, measured over 22 tumors (a subgroup of the 65 tumors), that are ordered vertically; each row corresponds to a sample. Blue color represents a low expression level and the red - high. The rows of the matrices A and B were not normalized, whereas those of D and E were normalized. Figures C and F are the distance matrices of the 22 tumors, where blue color represents a small distance and the red color represents large distance between two samples. Distances between samples are calculated in the 12 – dimensional space of the particular gene cluster used. This cluster of genes is one of several, generated by clustering the 1753 genes, based on all the 84 samples. The unique feature of this set of genes is the high similarity (high correlation) of their expression levels over these 22 tumors. Looking at image A one sees that some tumors have low (blue) and some others high (red) expression levels over nearly all the 12 genes. Clustering the samples using the expression matrix A and reordering the data accordingly generates the expression matrix B. Calculating the Euclidean distances between the 22 tumors (reordered as in B), generates the distance matrix C. Inspecting B, one sees that two subdivisions can be made; a big subgroup (of 15 tumors) shows low expression levels of the 12 genes, and the small subgroup (7 tumors) exhibits high expression levels. The same observation can be made observing the distance matrix C; the tumors in the big group are much closer to one another (blue) than to the other tumors. The tumors in the small group are far from the 15 tumors, but not that close to one another; the distance between them is represented mainly by green and yellow colors. The main point in this analysis is the use of a highly correlated set of genes to cluster a subset of samples. We can see that such a highly correlated group of genes holds important information, by means of which hidden structures in the data can be revealed. If the samples are also centered and normalized (as done by Perou et al), the data matrix A is replaced by D. The important information, that some samples had uniformly high, and others low expression over all 12 genes, is eliminated. When the samples are clustered according to D, the structures seen in the correspondingly ordered expression matrix E and distance matrix F are eliminated. This example demonstrates our assertion that by normalizing the samples before clustering one loses important information.

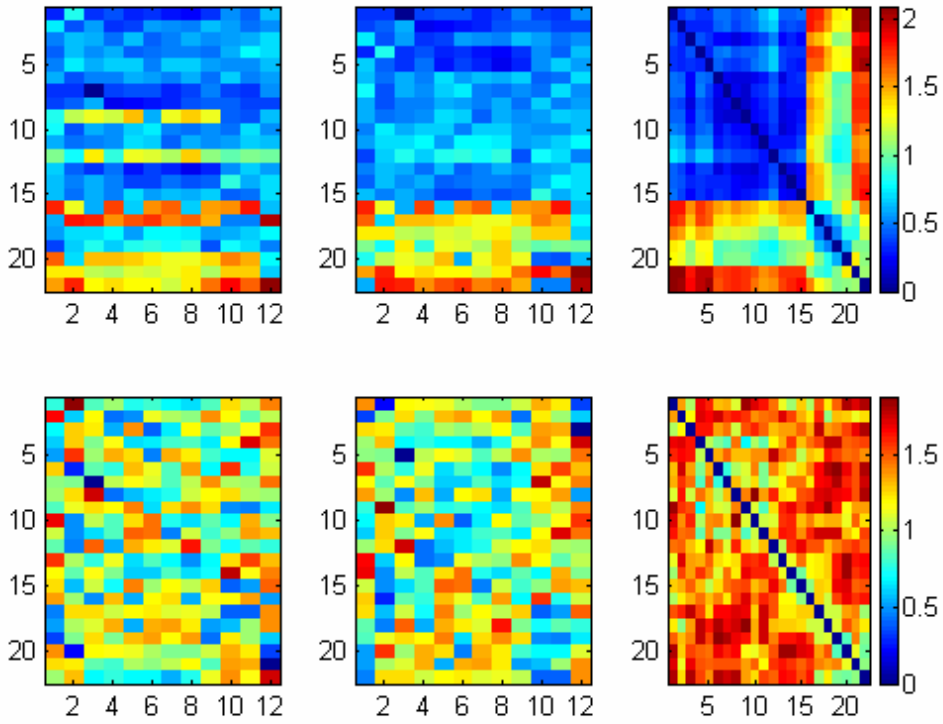

Figure 5.4. Figures A-B and D-E represent the expression levels of a cluster that contains 12 genes, in 22 tumors. Figures C and F are distances matrices of the 22 tumors. The rows in A-B were not normalized, and the rows in D-E were normalized. Looking on A-B one sees clear structures; tumors that show high expression level, and tumors that show low expression level of the genes. Looking on D-E one sees that after normalization the structure seen clearly in A-B has disappeared.

5.4.2.1 The clustering dilemma

All our clustering results were obtained using the Superparamagnetic Clustering algorithm (SPC) [12-13]. This is a stochastic algorithm, which maps the clustering problem onto a ferromagnetic spin model, which is simulated at a sequence of temperatures T. One measures the correlations between pairs of spins and uses their values to decide whether a pair of data points is to be assigned to the same cluster or not.

This decision is taken in two independent steps. In the first the values of the correlation G_{ij}, of the spins associated with data points i and j, are thresholded; if $G_{ij}>\theta$, the two points are assigned to the same cluster. This stage generates small, tight and highly reproducible clusters of points. The second stage is one of "growth": pairs of spins whose correlations did not pass the threshold are tested. Each data point is joined to the one with which it has the highest

correlation. Larger clusters are generated in this second stage, that grow out of the small clusters that were generated in the first step, which serve as the "cores" of the large ones. The second stage is based on comparing correlations of pairs which have low values of G_{ij}, and hence this stage is much more susceptible to statistical fluctuations and is less reproducible from run to run of the stochastic clustering process.

When we cluster a set of objects *O* (say a group of genes G1) on the basis of a set of features *F* (e.g. the expression level of the genes on a set of samples S1), we face a decision; should we use SPC *with* the growth step, or *without.* The first option will generate larger clusters, with a relatively small number of objects belonging to very small clusters. The second option will generate smaller clusters, and relatively many objects will belong to a "background" of very small clusters, of one or two members; the tradeoff is that the (non-background) clusters will be highly reproducible and represent "real" correlations, not some statistically un-significant effect of noise in the data.

The general strategy we adopted in this work was to use SPC *without growth* when we cluster genes; when clustering samples, we usually present the results of SPC *with growth,* unless otherwise stated. We will denote the operation of clustering *O* on the basis of the features *F* by either *O(F)* when the clustering operation is *without growth,* or as ***O(F)*** **(bold)** when SPC is used *with growth.*

5.4.3 Selecting a better set of genes

5.4.3.1 The 496 "intrinsic" genes

The goal of Perou et al was to develop a system for classifying tumors on the basis of their gene expression patterns. The full set of genes shown in Fig. 5.1 was not optimal for this purpose. Perou et al therefore turned to selecting their 496 intrinsic sets of genes to use as the basis for a new clustering analysis. The rationale behind this alternative gene subset was that specific features of a gene expression pattern, that are to be used to classify tumors, should be more similar for samples taken from the same tumor, and they should vary among different tumors. The 20 paired samples provided the basis data set on which a systematic search for such genes was performed. Perou et al initially ranked the selected 1753 genes according to such a similarity criterion, and decided to work with the top 496, their 'intrinsic' gene subset.

This set was obtained in a supervised way, using information that was not contained in the expression matrix; 1) that genes, which have the desired property, should be treated preferentially and 2). The decision to keep the top 496 genes (and not another number) may have been influenced by the quality of the resulting partition.

5.4.3.2 Analyzing the clusters obtained by CTWC

The main idea of CTWC is to identify subsets of genes and samples, such that when one of these is used to cluster the other, stable and significant partitions emerge [14] (see chapter 3).
The output of CTWC has two important components. First, it provides a broad list of gene and sample clusters. Second, for each cluster (of samples, say) I know which subset (of samples) was clustered to find it, and which features (genes) were used to represent it. My goal is to look for meaningful classifications of the tumors. I consider and test only clusters that passed the stability filter; if the resulting partition has stable components, I check weather it is biologically interesting.

Supervised search - Identifying genes that partition the samples according to a known classification.
This is a supervised test of clusters that were obtained in an unsupervised way. Denote by *C* a known classification of the samples; say into two classes, c_1 and c_2. Say I performed the clustering generation S(G), i.e., clustering samples of S, using their expression over the genes of G as the features. CTWC provides an easy way to rank the clusters of genes in *G* by their ability to separate the samples according to *C*. I evaluate for each cluster of samples S in *S* two scores, *purity* and *efficiency,* which reflect the extent to which assignment of the samples to c_1 corresponds to the classification *C*. These figures of merit are defined in equations (2).
Once a cluster S with high purity and efficiency with respect to c_1, say, has been found, I can trace back the desired pair of features and objects; the cluster (or clusters) of genes that were used as the feature set and the cluster of samples that were used as the object set to yield S. Hence by this method we identify the most natural group of genes that can be used to induce a desired classification.

Discovering new partitions

Another goal is to look for a new classification of the samples. For this I focus only on the most reliable clusters. The output of SPC provides a stability parameter by means of which I

can assess the reliability of each of the generated clusters. The higher the value of the stability parameters, the more reliable (statistically) is the cluster. Ranking the sample clusters according to their stability score enables to isolate the most reliable clusters, which have to be inspected more carefully. Useful hints for the meaning of such a cluster of samples may come from the identity of the cluster of genes, which was used to find it. More details about CTWC are reported in chapter 3.

In summary, there are two main differences between using the intrinsic set of 496 genes vs CTWC:

CTWC generates many highly correlated sets of genes. Some of these sets can be used to partition the samples in meaningful ways. In contrast, Perou et al used for their analysis only one (intrinsic) set of genes.

CTWC generates the set of genes to be used in an unsupervised way, in contrast to the way of selecting the 496 intrinsic genes.

In the results section I first show that the main observations reported by Perou et al were found also by CTWC. Second, that CTWC reveals new meaningful partitions.

5.5 RESULTS and DISCUSSION

5.5.1 Reproducing the results of Perou et al.

5.5.1.1 Clustering the genes using all the 84 samples, G1(S)

I clustered all 1753 genes used by Perou et al on the basis of their expression levels measured for the tumors and cell lines. In this process I used the same feature set (S = samples and cell lines) to cluster the same objects (G1 = 1753 genes) as Perou et al, and also used the same normalization. Hence this procedure compares directly the AVL and SPC methods.

I found several stable gene clusters. Seven of these were basically identical to gene clusters mentioned by Perou et al (they report nine, including their proliferation cluster). My analogues to their clusters h and i did not pass our criteria for stability, but I do see these two clusters also. In addition to these seven, we found additional stable gene clusters that were used in our subsequent CTWC analysis.

5.5.1.2 Clustering the 65 samples, using 1753 genes, S1(G1)

Even though I used different normalization and a different clustering algorithm, I agree with Perou et al that this operation does not yield interesting stable partitions of the samples.

5.5.1.3 Clustering the intrinsic genes vs G1(S)

By clustering their intrinsic gene set Perou et al identified four gene clusters, c, d, e, f of their Fig. 5.3. Of these, cluster d is identical to my gene cluster G21, and c has considerable overlap with G4.

5.5.1.4 Clustering the 65 samples; using the intrinsic set vs CTWC

The sample clusters (BLUE, GREEN, RED, YELLOW) that were found by Perou et al, using their intrinsic gene set, are found also by the CTWC procedure.

BLUE: The operation S1(G4) uses 10 genes that belong to my stable gene cluster G4 to cluster all 65 tumors. G4 is our homologue of cluster j of Perou et al. **S1(G4)** produces a sample cluster which is quite similar to the BLUE cluster of Perou et al; its members have high expression levels of G4.

GREEN: The operation **S1(G46)** produces a good homologue of the GREEN cluster of Perou et al. G46 is a cluster of 33 genes that are part of their proliferation cluster. Members of the GREEN cluster have low expression levels of G46 genes. Using G9 - a cluster of 13 genes (that are a subgroup of cluster g of Perou et al) - I can also separate their GREEN cluster from the other samples. Members of the GREEN cluster have high expression of the G9 genes.

RED: S1(G21) separates the members of the RED cluster from the other samples. G21 is homologous to gene cluster d of Fig. 5.3 of Perou et al, whose expression is high in the RED tumors.

YELLOW: This cluster is reproduced as one of the sample clusters obtained by the operation **S1(G4)** mentioned above. The expression level of the G4 genes on the YELLOW tumors is low.

I turn now to a detailed description of the results summarized above.

5.5.1.1 Clustering all Genes, using all Samples – G1(S), G1(S1)

The first step was to compare the structure that was found by Perou et al, shown in Fig. 5.1, to the corresponding features generated by SPC; by 1) clustering all the genes, **G1(S)**, using data from all the 84 samples (Fig. 5.1 b-j), and 2) clustering separately the 19 cell lines and the 65 tumors, based on all the 1753 genes. The clustering procedure was done by applying the SPC cluster algorithm [12-13] with and without the Directed Growth function (DG), as explain in the "clustering dilemma" section. The clusters that were generated by applying SPC without DG are very distinct and constitute the cores of the clusters that were generated by applying SPC with the DG function.

I first compared the gene clusters generated by SPC to those generated by AVL that are presented in fig. 5.1 of Perou et al (c to j and the proliferation cluster (prl) that is reported in the supplementary information) (see also Table 5.2). Clustering the full set of genes, using SPC with DG generated the dendrogram presented in Fig. 5.5. The genes that belong to the clusters of Fig. 5.1c-j, and the "proliferation" (prl) cluster generated by Perou et al are labeled and marked on the SPC dendrogram (Fig. 5.5). In this figure one sees that the clusters marked as interesting by Perou et al were also found by SPC. In order to put this statement on a

quantitative ground, I scanned all the nodes (clusters) of the SPC dendrogram, searching, among the SPC clusters for its homologue in Fig. 5.1 (Perou et al). The homologue SPC cluster is defined as the one with the highest purity (P) and efficiency (E); denoting by X a cluster of Fig. 5.1 and by Y a cluster of SPC,

$$E = \frac{|X \cap Y|}{|X|}, \quad P = \frac{|X \cap Y|}{|Y|}. \tag{3}$$

The results of this search are reported below, in Table 5.1 and Fig. 5.5, and the blacks arrows in the dendrogram of Fig.5 points to the optimal homologoues boxes (clusters). The results obtained by AVL and by SPC using DG are compared in column 1 of Table 5.1. When the data were clustered by SPC without DG only the cores of the clusters were found, hence, higher purity but lower efficiency are expected. This is indeed the case; as seen in column 2, all but one of the Perou et al clusters were found with purity 1 and somewhat lower efficiency than in column 1. Only cluster h (Fig. 5.1h) was not found by SPC without DG.

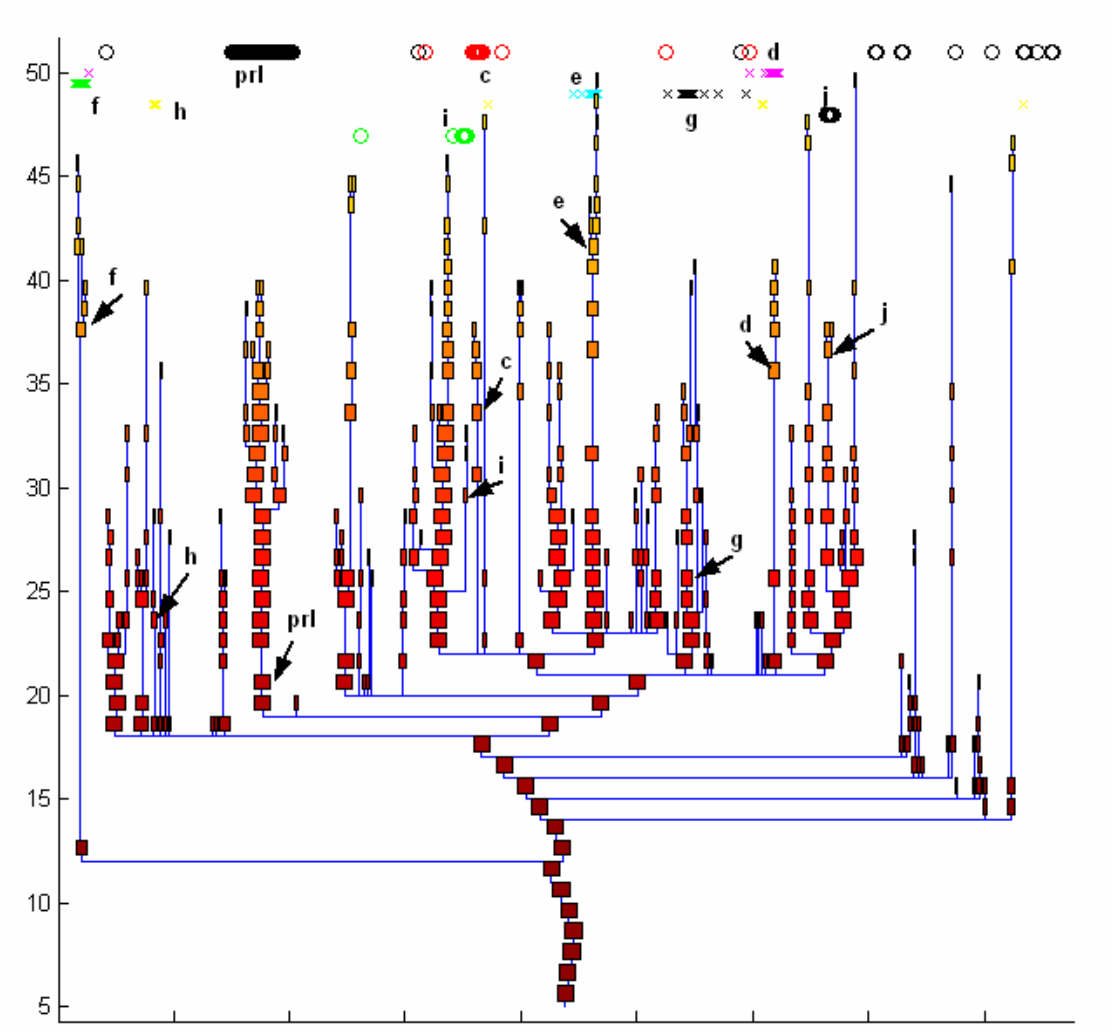

Figure 5.5. G1(S) - The dendrogram generated by applying SPC on the entire set of 1753 genes based on the 84 samples. The genes are ordered according to the dendrogram. Those that belong to one of the Peoru at al clusters of Fig.1 were identified and marked as follows: cluster c; red o, cluster d; purple x, cluster e; blue x, cluster f; green x, cluster g; black x, cluster h; yellow x, cluster i; green o, cluster j; black o. The blacks arrows mark the optimal homologues among the SPC clusters. As one sees from the visual comparison above, except for several outlier genes, most of the clusters of Perou et al were found also by SPC, with high levels of purity and efficiency.

Column 4 of Table 5.1 presents comparison of the Perou et al clusters to those generated by the first step of CTWC using SPC without DG. CTWC records only clusters that passed a minimal stability threshold, defined externally, as shown in the dendrogram Fig. 5.6. I denoted by GX clusters of genes, where X is the number of the cluster assigned by CTWC. The results of the CTWC algorithm corresponding to the use of G1(S) as the data set to be its starting point (all the 1753 genes based on the 84 samples),are reported in the web site: **http://www.weizmann.ac.il/home/cskela/perou/perou_all_samples2/ctwc_all_samples2.html.**

CLUSTERS FROM FIGURE 1	Running SPC With Directed Growth	Running SPC Without Directed Growth	First iteration of CTWC (SPC) Without Directed Growth
Cluster c	P=0.72 E=0.72	P=1 E=0.5	G11 P=0.9 E=0.55
Cluster d	P=0.86 E=0.82	P=1 E=0.78	G16 P=1 E=0.78
Cluster e	P=0.76 E=0.88	**P=1 E=0.88**	G12 P=1 E=0.88
Cluster f	P=0.66 E=1	P=1 E=0.85	G23 P=0.76 E=0.92
Cluster g	P=0.86 E=0.73	P=1 E=0.7	G15 P=0.93 E=0.55
Cluster h	P=0.8 E=0.6	not found	Not found
Cluster I	P=0.89 E=0.81	P=1 E=0.72	Not found as a stable cluster
Cluster j	P=0.7 E=1	P=1 E=0.9	G4 P=1 E=0.9
Prl (proliferation genes)	P=0.8 E=0.8	P=1 E=0.45	G17 P=1 E=0.4

Table 5.1. Comparison of Perou's clusters to the clusters generated by SPC. The clusters in Fig.5 were scored for homologies with the clusters of Fig. 5.1 (Perou et al) and the most homologous ones were depicted by arrows in Fig. 5.5 and analyses here for purity and efficiency. Column 4 shows the clusters found by CTWC (see Fig. 5.6a) as the corresponding to these in column 1.

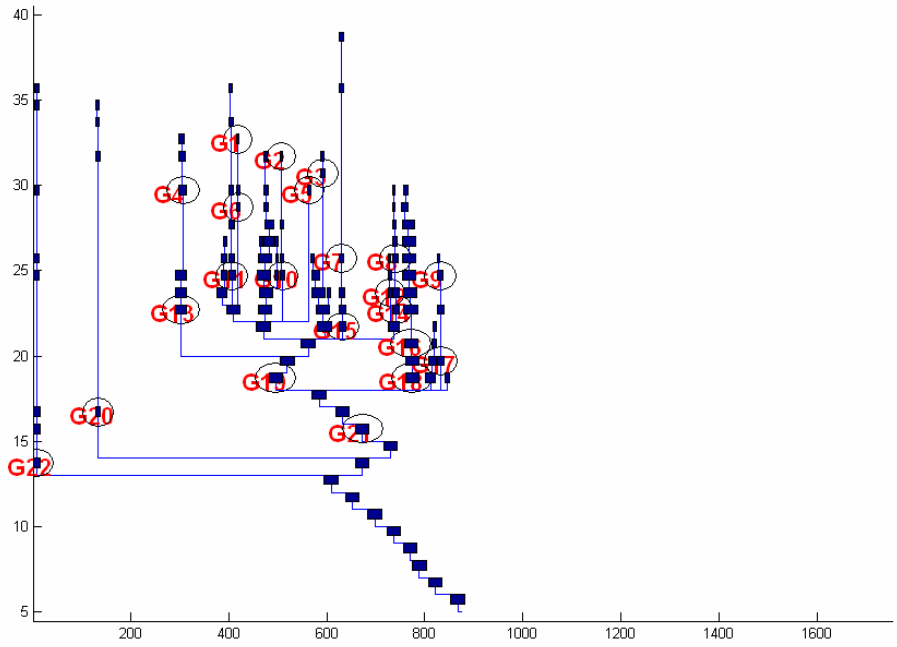

Figure 5.6a. G1(S) : SPC dendrogram with stable clusters. Clustering the set of all genes using all 84 samples, by SPC without DG, generates the dendrogram presented above. Only the clusters that were found by CTWC as stable are marked in the dendrogram.

I clustered the genes also on the basis of their expression levels in the 65 tumors, G1(S1). This step is the natural starting point for the CTWC procedure. The resulting dendrogram, Fig. 5.6b, shows the stable gene clusters obtained by this procedure.

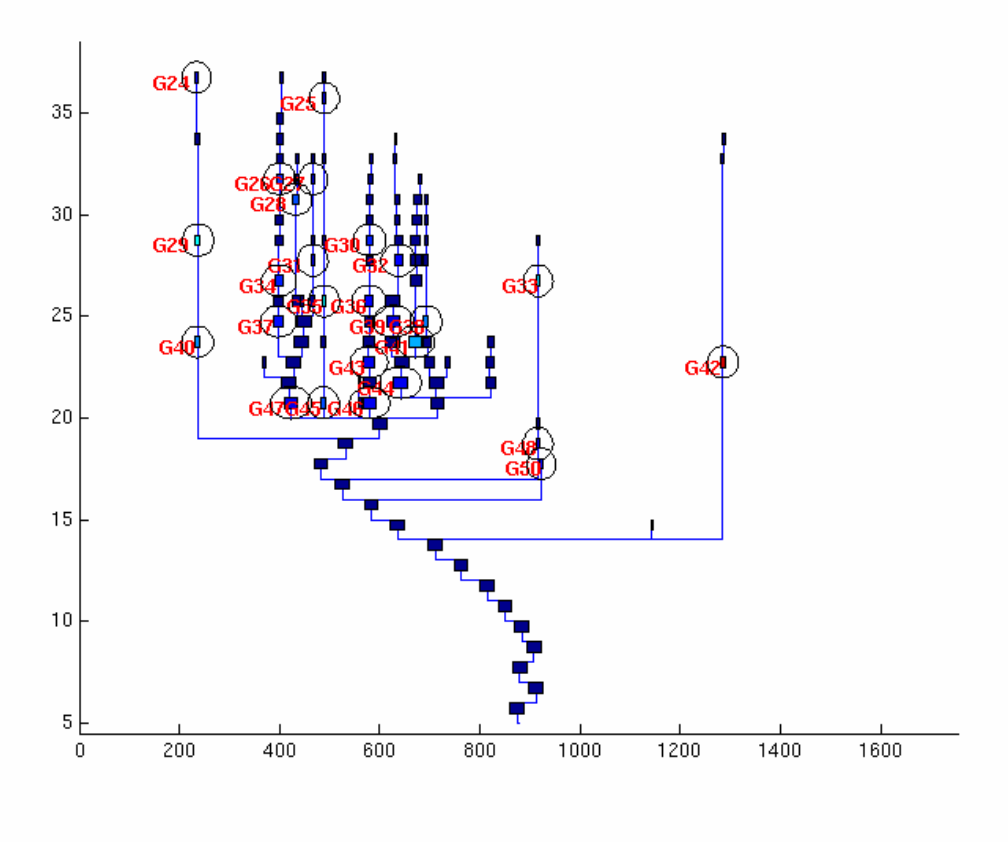

Figure 5.6b. G1(S1): SPC dendrogram showing only stable clusters. Clustering the set of all genes using the 65 tumors, by SPC without DG, generates the dendrogram presented above. Only the clusters that were found by CTWC as stable are marked on the dendrogram.

Since I normalized the genes in the same way as Perou et al, the results presented here (Table 5.1 column 4) indicate that most clusters that were found by Perou et al are also found by SPC. Two of these (h and I, table 5.1) were not identified by SPC as stable and significant. On the other hand, SPC did find several additional stable gene clusters. The dendrogram presented in Fig. 5.6a represents the 23 gene clusters that passed the stability threshold of CTWC. As can be seen in column 4, 7 out of the 23 CTWC clusters were also identified by Perou et al as significant clusters. SPC found additional 16 sets of genes as stable clusters. Another 3 sets of genes (G30, G46, G35, Fig. 5.6b) that divide the samples to significant subgroups will be discussed later. These clusters were generated after clustering all the 1753 genes (G1), using the 65 tumors (Fig. 5.6b). More details are reported in the site **http://www.weizmann.ac.il/home/cskela/perou/perou_last_analysis/rows/G1_PerouCtwc_S1.html**).

The sets of genes that are shared between the clusters in Fig. 5.1 and in Fig. 5.6a (column 1 and 4 respectively of Table 5.1) are listed and classified. The number of genes in each cluster is given in parentheses in column 1. Additional clusters derived from Fig. 5.6a and 6b (that are not homologous to clusters in Fig. 5.1).

Stable CTWC clusters	Perou et al	Remarks
G11 (11)	Cluster c	Genes which are expressed by Endothelial (blood vessel) cells.
G16 (18)	Cluster d	Genes which are expressed by stromal/Fibroblast cells
G12 (16)	Cluster e	Genes that are expressed by basal epithelial cells: less differentiated cells that undergo continuous mitosis.
G23 (23)	Cluster f	B-lymphocytes genes.
G15 (16)	Cluster g	Contains cluster G9.
G4 (10)	**Cluster j**	**This set of genes contains the Estrogen receptor and to 3 others transcription factors; GATA-binding protein 3, X-box binding protein 1 and Hepatocyte factor 3a, that have a relation to the Estrogen Receptor pathway.**
G17 (43)	Cluster prl	Genes, which correlate with cellular proliferation rate.
G21 (12)	-	**This cluster contains the Erb-B2 gene and several other genes, many of them located in the Erb_B2 amplicon.**
G9 (13)	-	**Cluster contains genes characteristic of the normal breast epitheliumium.**
G46 (33)	-	**Proliferation genes; part of cluster G17 (33/43).**
G30 (15)	-	**Proliferation genes. part of cluster G46 (15/33).**
G35 (11)	-	**B-lymphocytes genes. part of cluster G23 (11/17)**

Table 5.2. Clusters of genes that were obtained by Perou et al (Fig. 5.1) as compared with CTWC (Fig. 5.6a and b). The sets of genes that are shared between the clusters in Fig. 5.1 and in Fig. 5.6a (column 1 and 4 respectively of Table 5.1) are listed and classified. The number of genes in each cluster is given in parentheses in column 1. Additional clusters derived from Fig. 5.6a and 6b (that are not homologous to clusters in Fig. 5.1) are also included here. The clusters shown in bold letters further analyzed later. The detailed lists of the genes in each cluster are given in the appendix.

5.5.1.2 Clustering the 65 samples, using all genes - S1(G1)

Clustering the cell lines and the tumors, separately, based on all the 1753 genes (Perou et al) generated the dendrograms of Fig. 5.1a. They normalized also the samples (before clustering), according to equation (1) of the last section. I used SPC without DG, for the clustering of the 65 tissues without this normalization, using all 1753 genes, as shown in the dendrogram of Fig. 5.7. Similar analysis with DG is given in Fig. 5.8. The main difference between my analysis and that of Perou et al at this stage is the normalization. As explained in the "normalization dilemma" in section 5.4.2.1, normalization before clustering can change dramatically the final results. The example shown in the "normalization dilemma" section shows clearly why we prefer not to normalize the samples. Figure 5.9 presents the two distance matrices of the 65 tumors; in 9A the samples are ordered according to the dendrogram obtained with DG, Fig. 5.8, using SPC without normalization, whereas in 9B the samples are ordered according to the dendrogram of Fig. 5.1a of Perou et al. No clear structures can be seen in 9B; this is because the normalization eliminates some of the important information contained in the expression levels of the 1753 genes. Clearer structures can be seen in 9A; although the distance matrix 9A highlights interesting partitions, that are not seen in 9B, I came to the same conclusion as Perou et al: namely, that the entire set of genes is not suitable to classify the tumors and I should characterize them using different subsets of genes.

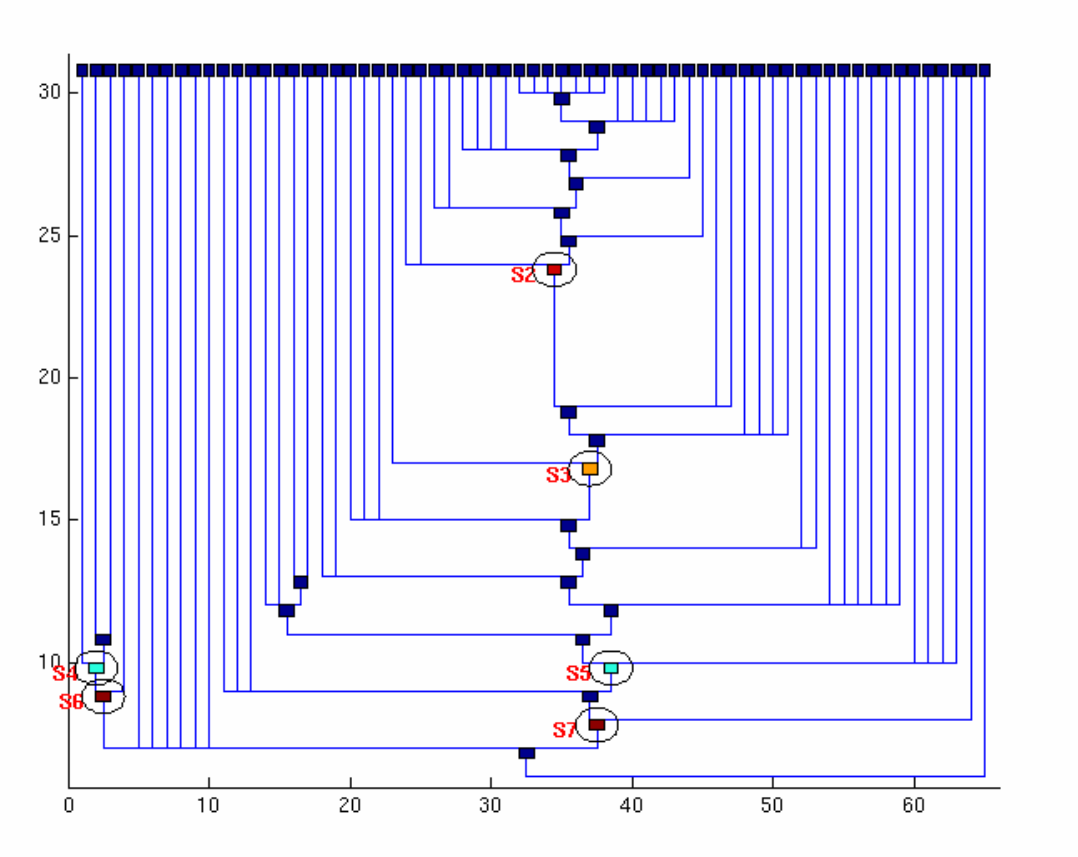

Figure 5.7. SPC dendrogram; generated after clustering the 65 tumors based on all 1753 genes, by using SPC without DG. Seven clusters passed the stability threshold, and are marked by circles on the dendogram.

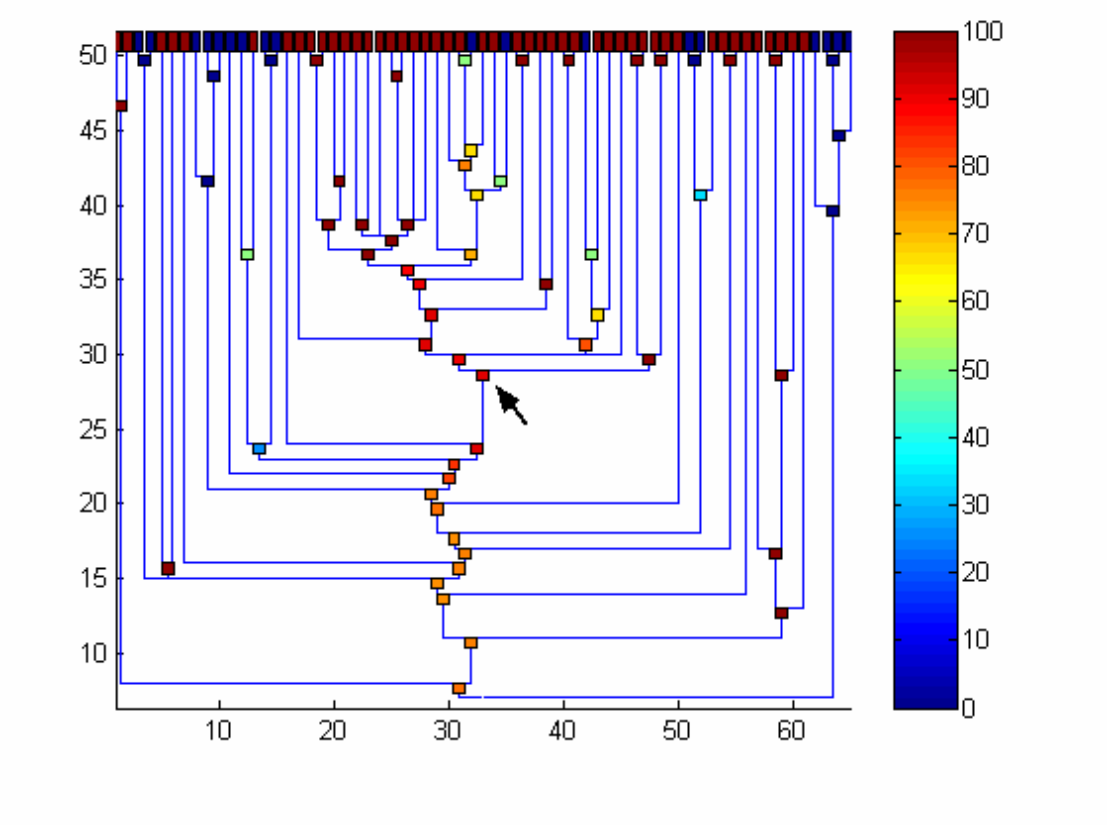

Figure 5.8. The dendrogram generated by SPC after clustering the 65 tissues based on the entire set of 1753 genes without normalizing the samples and with Directed Growth. The tumors in the cluster indicated by the black arrow are separated from the others due to their high correlation. This cluster is also seen clearly in the distance matrix of Fig. 5.9A. Boxes represent the clusters; each box is colored according to the percentage of ER+ tumors (tumors that contain the estrogen receptor protein) as indicated on the colorbar on the right hand side.

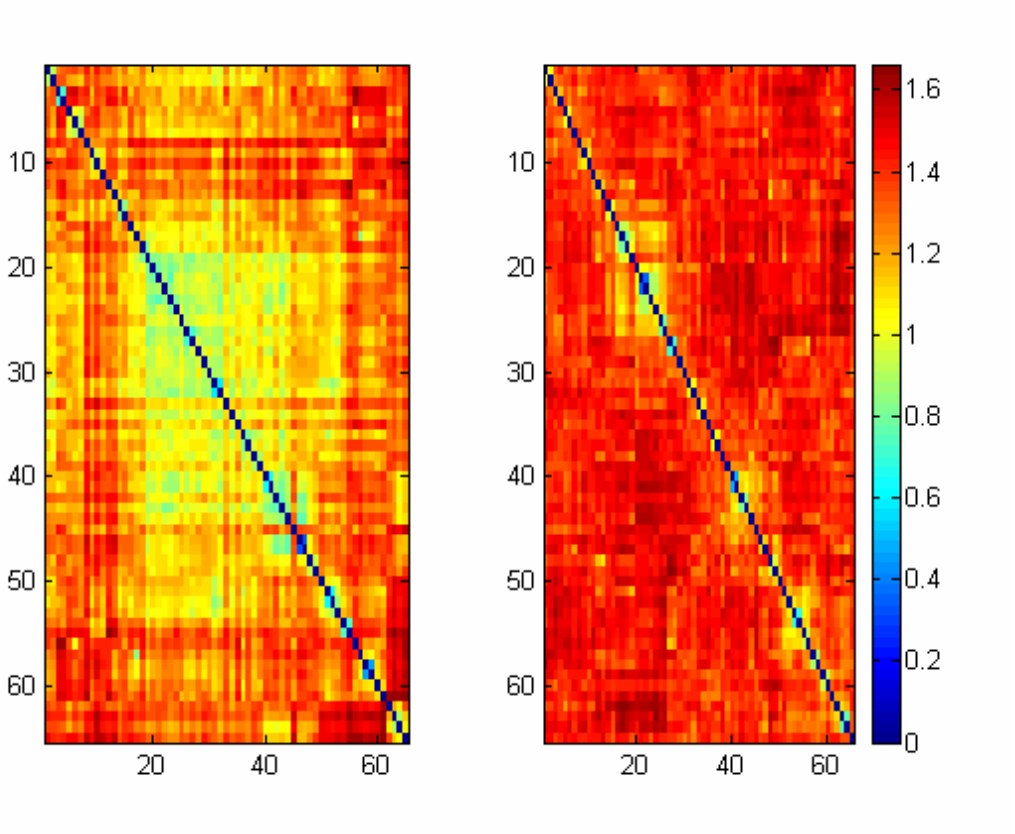

Figure 5.9. Two distance matrices of the 65 samples. A; using SPC (with DG) to cluster the samples, without normalizing them, according to Fig. 5.8. B; using AVL to cluster the (normalized) samples, according to Fig. 5.1. The tumors in both distance matrices are ordered according to the corresponding dendrogram. In B no significant partitions are seen, in contrast to A, where a group of tumors (in the middle) show a significant correlation among themselves, compared to the other samples. This group of samples is marked in the corresponding dendrogram of Fig. 5.8 by the black arrow.

5.5.1.3 The gene clusters obtained by Perou et al from the intrinsic set

Perou et al highlight four gene clusters obtained by clustering their intrinsic set of 496 genes. Of these, cluster d is almost identical to our cluster G21, found by G1(S) (without DG, see website http://www.weizmann.ac.il/home/cskela/perou/perou_all_samples2/rows/r1_ctwc_all_samples2_c1.html). Their cluster c has many genes, including the estrogen receptor and the gata-binding protein 3, in common with our cluster G4, also found by G1(S). The other two clusters, e and f, were not found in the G1(S1) and G1(S) procedures as stable clusters.

5.5.1.4 Clustering the 65 samples by CTWC compared to the intrinsic genes of Perou et al

In this section I compare the results obtained by Perou et al by using the set of 496 "intrinsic" genes, to those of the CTWC algorithm. I used the supervised test described in the method section 5.4.3.2 to look for CTWC clusters of genes that have the ability to separate the samples according to the classifications presented in Fig. 5.3a.

To represent and evaluate the classifications reported by Perou et al in their Fig. 5.3, I used the following; "BLUE cluster"; includes the 36 samples colored in blue in Fig. 5.3. Similarly GREEN, RED, YELLOW clusters refer to the samples colored accordingly in Fig. 5.3. I call 'ER+' the 46 tumors whose cells contain the Estrogen Receptor protein according to the supplementary information of Perou et al[1]. Eleven tumors do not contain ER protein are called "ER-"; their cells do not contain the Estrogen Receptor protein.

The definition ER+ or ER- is according to the clinical estrogen receptor status, that can be determined by several techniques. For example by immunohistochemistry, using antibodies to ER.

Our full results, obtained by CTWC, using the G1(S1) as the data to be the starting point (all genes based on the 65 tumors), are presented in the site:
http://www.weizmann.ac.il/home/cskela/perou/perou_last_analysis/PerouCtwc.html.

The BLUE cluster: The tumors contained in this cluster are characterized by a relatively high expression level of many genes that are known to be expressed by breast luminal[2] cells (Fig. 5.1j). 32 of the 36 tumors in the BLUE cluster are ER+; hence if this cluster is to be used as a classifier for ER+, it has purity=0.88. Of the remaining 4 tumors, 2 tumors are ER- (NORWAY 48, BE and AF), and the others were not tested for the ER clinical status. However the entire collection contain 48 ER tumors. Since the BLUE contain only 32 of the 48 ER+ tumors; it has efficiency = 0.66. The other 16 ER+ tumors fell within the Green, Red or Yellow clusters. I set out to find, weather using CTWC, I can partition the samples according to theirs ER+/ER- classification, in an unsupervised way. I found two clusters of genes, which achieve this partition; the first, G4, has considerable overlap with cluster c of

[1] (http://genome-www.stanford.edu/breast_cancer/molecularportraits/SITable3_new.html).

[2] Cells, those are located in the lumen surface.

Fig.3 of Perou et al, and is related to the estrogen receptor pathway. The second set of genes, G30, is related to cell proliferation.
Now I will cluster the 65 tumors (S1) using these G4 and G30.

Clustering S1 using G4; **S1(G4):** Clustering all the 65 tumors (S1) using the expression levels of the 10 genes contained in G4 (see Table 5.3 and apendix) generated the dendrogram presented in Fig. 5.10A. The variation in the expression of this cluster correlated well with the direct clinical measurements of the ER protein levels in the tumors (Perou et al, supplementary information). G4 is practically identical to cluster j of Fig. 5.1 and cluster c in Fig. 5.3 (Perou et al).

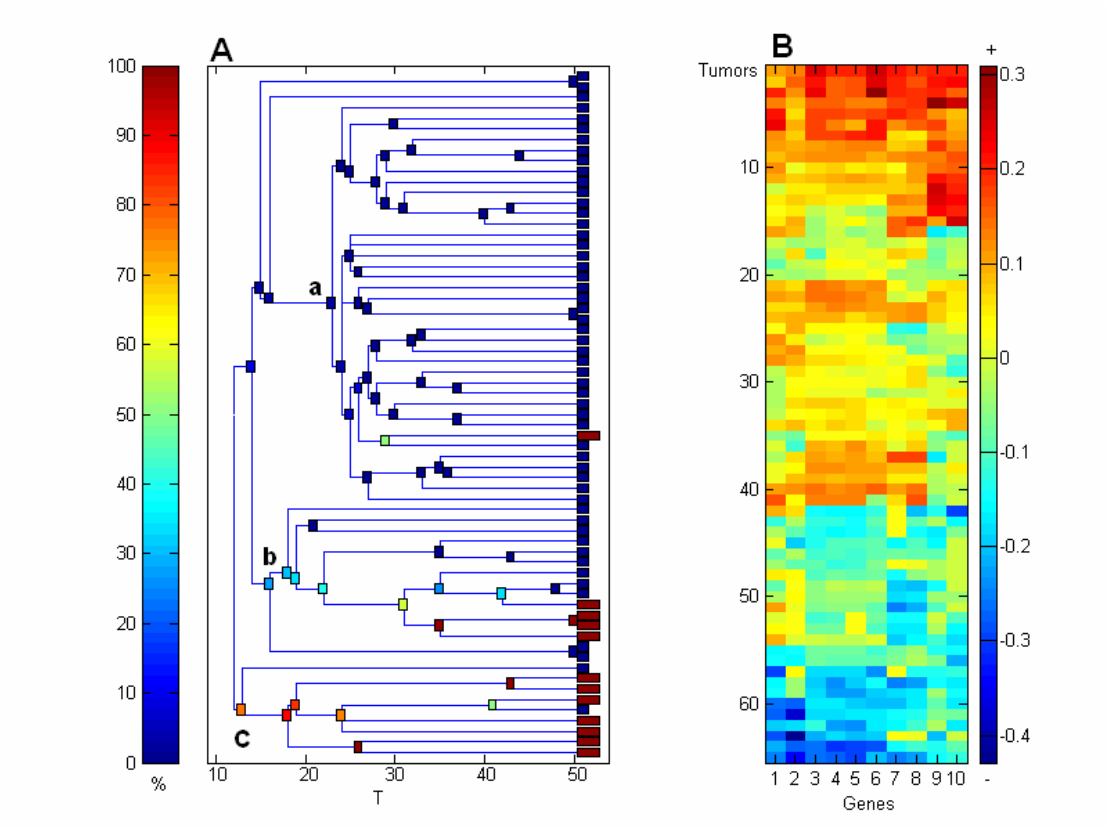

Figure 5.10. S1(G4): clustering the 65 tumors using the expression levels of gene cluster G4 yields three clusters; a contains most of the tumors of the BLUE cluster. The boxes that represent subclusters are colored according to their percentage of ER- tumors (see colorbar on left).

In the dendrogram Fig. 5.10A sample clusters appear as boxes. These boxes were colored according to purity, i.e. the percentage of ER- samples, ranging from 100% (red) to 0% (i.e. 100% ER+) (blue) as shown on the color bar on the left. In Fig. 5.10B, the samples are ordered according to the dendrogram Fig. 5.10A. The colors (see bar on the right hand side) represent the expression levels of the 10 genes, with red denoting high and blue low values.

SPC generated 3 main branches (clusters); the upper of which is shown as a, the middle as b, and the lowest as c. Cluster a, the biggest (41 samples) exhibits high expression level of G4 and contains most of the samples from the BLUE cluster (34 out of 36). The two missing tumors are NORWAY 48 (BE and AF), and they were indeed ER-. Cluster a has purity P = 0.82 and efficiency E = 0.94 with respect to the BLUE cluster (See results in Table 5.3). Cluster **a** includes only one ER- tumor (NORWAY 65) and four tumors, that belong to the GREEN cluster; The 3 NORMAL samples and the two NORWAY 112 (which are ER+) and the STANFORD 37 (which was not tested for ER). The ability of G4 to separate ER+ from ER-will be discussed in Section 5.5.2.1. Cluster **b** (15 samples) is characterized by lower expression level of G4. Eleven samples from this cluster are ER+ and the others ER-. This cluster includes part of the GREEN cluster and most of the RED cluster. Seven of the nine tumors of cluster c are ER- and this cluster
coincides almost exactly with the YELLOW cluster.

S1 (G4) With Growth (Fig.10)	**BLUE (36)**	**GREEN (7)**	Red (7)	Yellow (8)
Cluster **a** (41)	34 P=82 E=0.94	3	1	0
Cluster **b** (15)	2	4	5	0
Cluster **c** (9)	0	0	1	8 P=1 E=0.9

Table 5.3. Three clusters emerged by clustering S1 based on G4 (Fig. 5.10). P is Purity and E is Efficiency. The numbers in the table indicate how many samples of each cluster of Perou et al are in our clusters a, b, c. The numbers in parentheses are the number of tumor in the group.

More details on the **S1/G4** analysis are reported in the site:
http://www.weizmann.ac.il/home/cskela/perou/perou2_ctwc_all/c1r4_out.html

Clustering S1 using G30; **S1(G30)**: Clustering S1 based on the cluster G30 (see Table 5.2) generated the features shown in Fig. 5.11. G30 contains 15 genes that are related to cell proliferation. Tumors that show high expression levels of these genes may have a high rate of cell cycle proliferation and may be more aggressive. The dendrogram in Fig. 5.11A shows

clusters colored according to their content of the ER- group, ranging from red (highest) to blue (lowest). Clusters a1 contains a high proportion of ER- tumors (see results in Table 5.4). The tumors of clusters **a** and **b** exhibit high expression levels of G30 genes, as seen in the expression matrix of Fig. 5.11B. Most tumors contained in these two clusters are clinically characterized as ErbB2- (see Table 5.4). Clusters **c**, **b** have high proportions of ER+ tumors. Cluster **c**, which exhibits medium expression level of G30 genes, is the best "ER+ classifier"; it contains most of the ER+ tumors (P=0.86, E=0.79). Cluster **b** contains 'special' ER+ tumors that have relatively high expression levels of the G30 genes. These tumors are clinically characterized as ErbB2-.

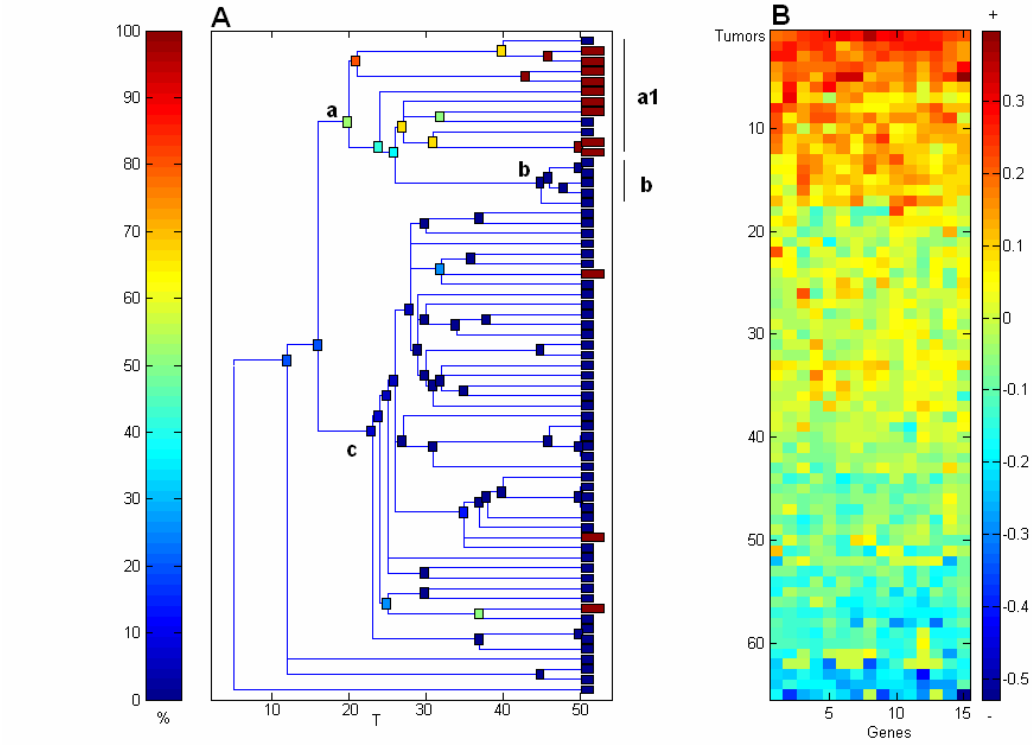

Figure 5.11. S1(G30): Cluster c has fairly high overlap with the BLUE cluster and contains mainly ER+ tumors. The boxes are colored according to the percentage of ER- tumors.

It may be concluded that the proliferation rate of most ER+ tumors is lower than that of the ER- tumors, and tumors that are clinically characterized as ER-/ErbB2- (the proteins ER and ErbB2 are not expressed) show the highest proliferation rate and may be the most aggressive (see cluster a in Table 5.4).

More details on the **S1(G30)** analysis are reported in the site:

http://www.weizmann.ac.il/home/cskela/perou/perou2_ctwc_all/c1r30_out.html

The clustering of S1 according to G4 used the ER among the genes of G4 and it is likely that ER+ samples will be partitioned. On the other the clustering on the basis of G30 used genes,

which correlated with proliferation and nevertheless partitioned the ER+ tumors as a cluster. This indicates correlation between ER and proliferation.

S1/G30	Blue (36)	Green (7)	Red (7)	Yellow (8)
Cluster a1 (11)	**5**	0	1	5
Cluster b (5)	3	0	0	2
Cluster c (44)	P=0.63 E=0.77	7	5	1

Table 5.4. Three clusters emerged by clustering S1 based on G30 (Fig. 5.11). P is Purity and E is Efficiency. The numbers in the table indicate how many samples of each cluster of Perou et al are in our clusters a1, b, c. The numbers in parentheses are the number of tumor in the group.

The GREEN cluster (normal breast like). The Green cluster (see Fig. 5.3) is defined by Perou et al as 'normal breast-like', includes the three normal breast samples. These samples are characterized by high expression levels of genes expressed by basal epithelial cells that grow in normal tissues (Fig. 5.1e).

Clustering S1 using G46; **S1(G46):** My best homologue to the GREEN, normal breast-like, cluster was achieved by clustering S1 based on genes of cluster G46 (see list in appendix). This cluster contains 33 genes whose levels of expression correlate with cellular proliferation rates. In the dendrogram Fig. 5.12A the clusters are colored according to their proportion of the GREEN group (Fig. 5.3). Cluster a (see Fig. 5.12) includes most of the green tumors; *Purity* = 0.7 (9/13), *Efficiency* = 0.8 (9/11). Looking at the expression matrix Fig. 5.12B, one sees that the expression levels of the G46 genes are lowest in the tumors contained in cluster **a**. I may assume that these tumors are growing slowly and, therefore, may be less aggressive. In contrast, clusters b, c and d show high expression levels of the G46 genes, therefore, may be more aggressive. The tumors of cluster **d** exhibit the highest expression levels of the set G46, as can be seen on the expression matrix. This cluster contains the ER-/ErbB2- tumors, which leads to the assumption that the absence of these two proteins can be markers for aggressiveness. The correlation of this with clinical studies should be explored. The tumors contained in clusters b and c are characterized by ER+/ErbB2-, ER-/ErbB+ or ER-/ErbB2-. In

other words, all these tumors are missing at least one of the two ER, ErbB2 proteins, which may be the reason for being also aggressive.

The red arrows point to the locations of the 3 'responder' tumors before the doxorubicin treatment, and the black arrows point to the same tumors after the treatment. According to the expression matrix, as a result of the treatment the proliferation rate of the cells changed. These observations may have interesting clinical implications. One may hypothesize that this set of genes may be used as predictors for the level of response to doxorubicin treatment. It may be possible to suggest that: tumors that belong to cluster e, and exhibit intermediate medium expression levels, have higher chances to respond to doxorubicin treatment.

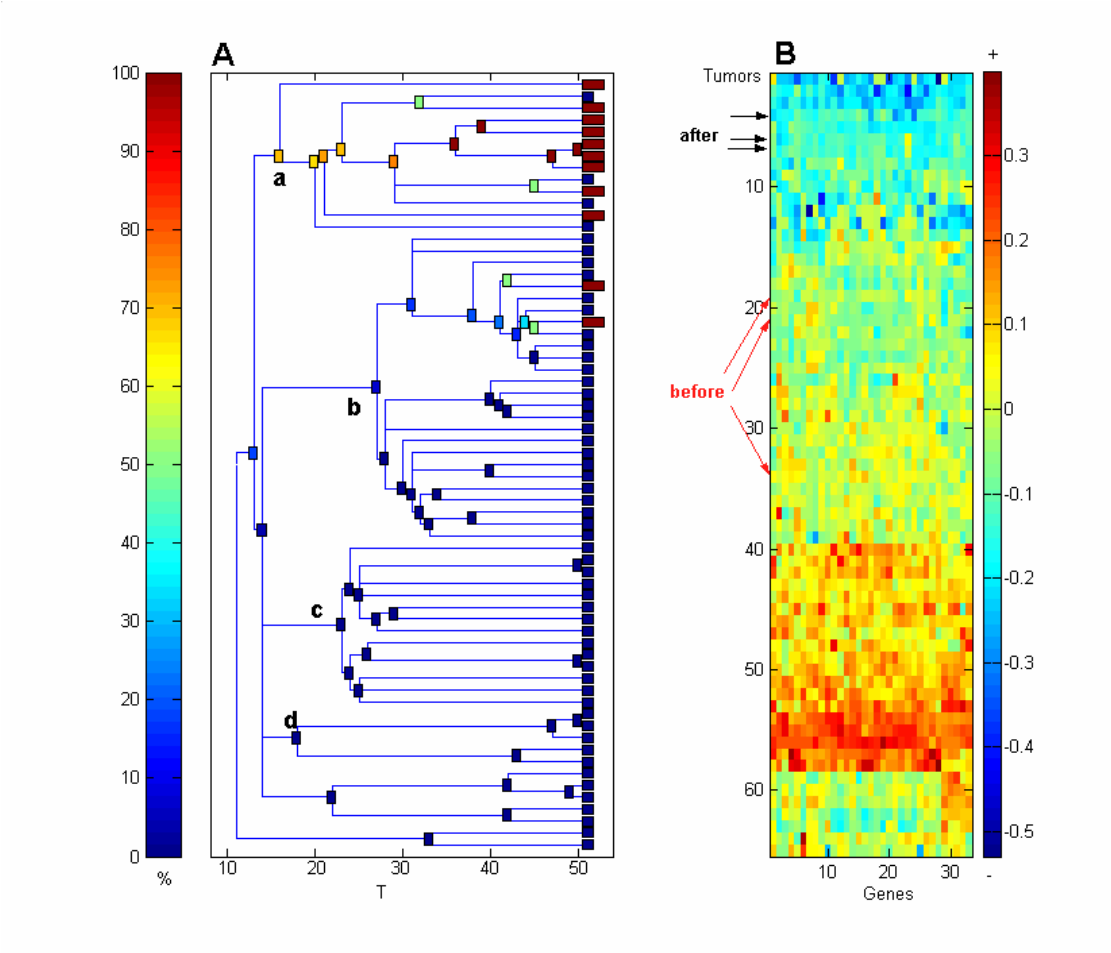

Figure 5.12. S1(G46) High, intermediate and low levels of proliferation related genes. The boxes are colored according to the percentage of the GREEN cluster; a is homologous to the GREEN cluster. The arrows indicate the three tumors that responded to the Doxorubicin treatment, before and after (see discussion in the New Results section).

http://www.weizmann.ac.il/home/cskela/perou/perou2_ctwc_all/c1r46_out.html.

Clustering S1 using G9; **S1(G9):** Another set of genes (G9) succeeded to separate the GREEN cluster from the others. The G9 genes are part of a bigger cluster, shown in Fig. 5.1g. This gene cluster, defined by Perou et al as 'adipose-enriched/normal breast genes, are highly expressed by normal breast epithelium cells. The clusters shown in the dendrogram Fig. 5.13A are colored according to their proportion of the Green group. As one sees in Fig. 5.13A, most of the GREEN tumors (8 out of 11) were clustered (cluster **c**) separately from the others (clusters **a** and **b**). These GREEN tumors exhibit higher expression levels of the G9 genes, according to the expression matrix shown in Fig. 5.13C. Most of the ErbB2- tumors were clustered together in cluster **a** separately from the GREEN tumors, and exhibit lower expression levels of the G9 genes (not shown).

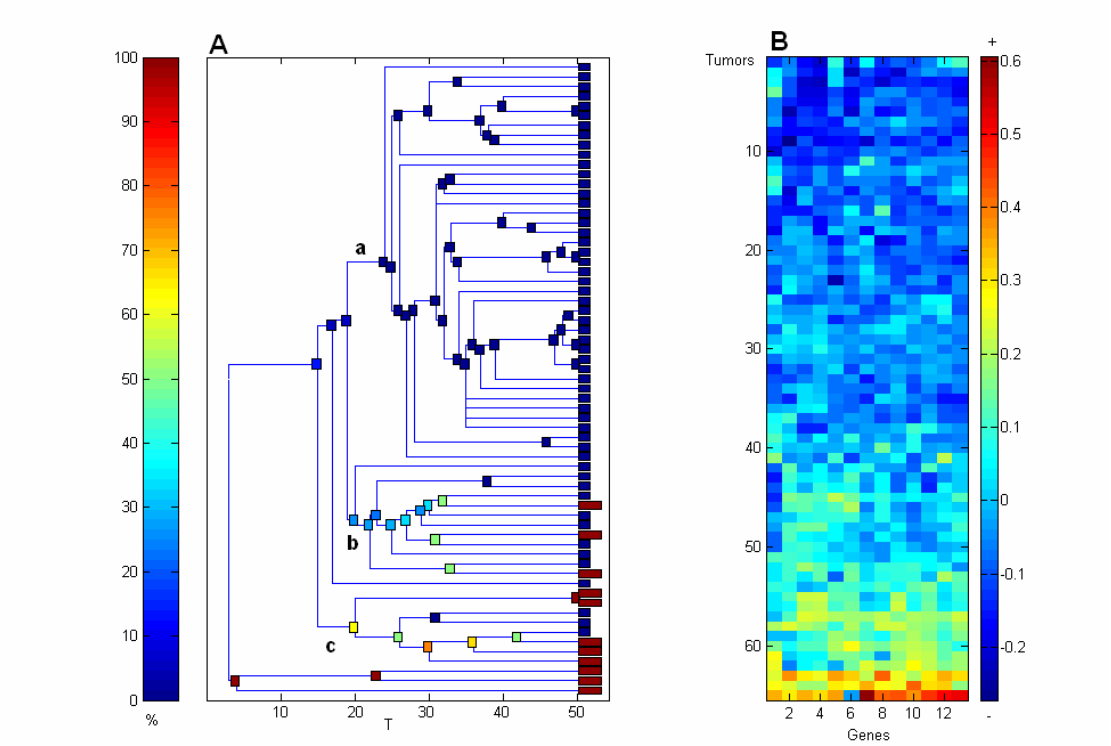

Figure 5.13. S1(G9) also yields a homologue of the GREEN cluster, which is denoted by c. The members of the GREEN cluster are colored in red.

The RED cluster (ErbB2 over expression). This cluster contains tumors that were characterized by high expression level of a subset of genes that contain the ErbB2 oncogene (these genes are contained in cluster d of Fig. 5.3).

Clustering S1 using G21; **S1(G21):** The best cluster of tumors homologous to the RED cluster was found by clustering S1 based on gene cluster G21 (see list in appendix). Cluster G21 is homologous to the gene cluster d of Fig. 5.3. In the dendrogram presented in Fig. 5.14A the clusters are colored according to their proportion of the RED samples. The RED samples are represented by red boxes and show high expression of ErbB2 (cluster **c**). Looking at the

expression matrix in Fig. 5.14B, one sees that theses tumors exhibit high expression levels of the G21 genes. Although these tumors do not form a single stable cluster, they are well separated from most of the other tumors, that form one big cluster, marked by **a**. This cluster divides to two sub-clusters. Most of the tumors in sub-cluster **b** are ER+ and ErbB2-; *Purity* = 0.86 (31/36), *Efficiency* = 0.64 (31/48). Very low expression levels of the G21 genes characterize this cluster.

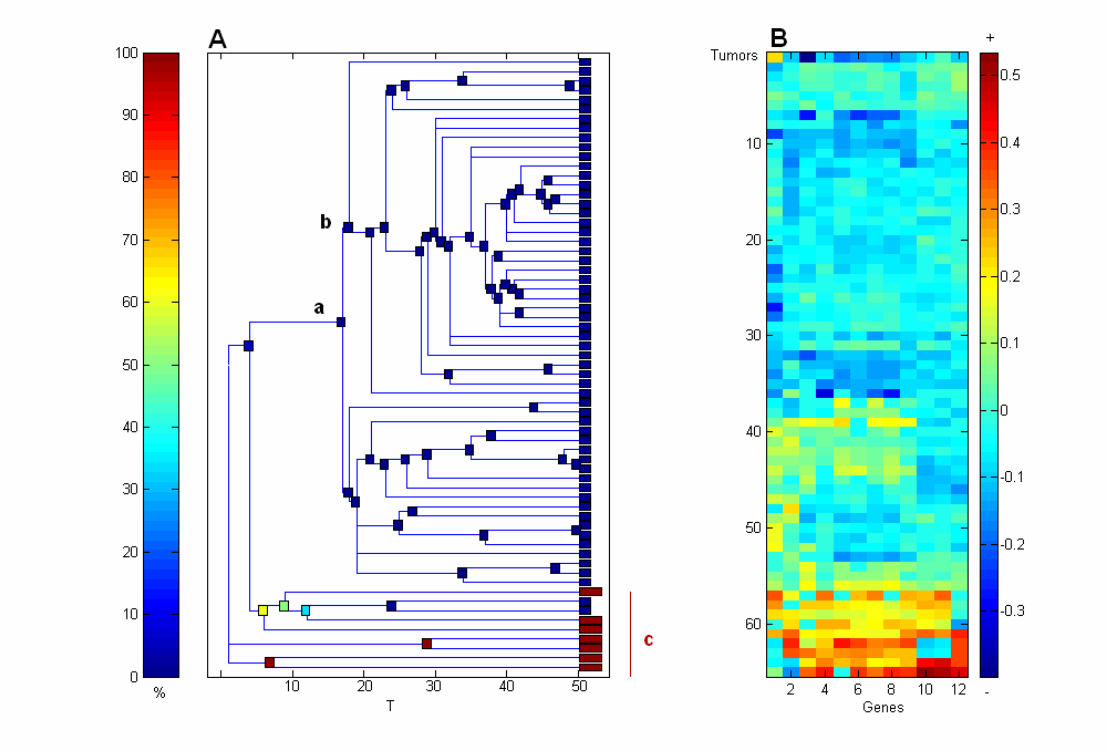

Figure 5.14. S1(G21); the group of tumors c shows high expression levels of G21, and belong to the RED cluster (red boxes).

More details about the **S1(G21)** analysis are reported in the site

http://www.weizmann.ac.il/home/cskela/perou/perou2_ctwc_all/c1r21a_out.html

The YELLOW cluster. This cluster is recaptured nearly perfectly as cluster c of the **S1(G4)** procedure (see Table 5.2 and Fig. 5.10). See site:

http://www.weizmann.ac.il/home/cskela/perou/perou2_ctwc_all/c1r4_out.html

5.5.1.5. The gene clusters obtained by Perou et al from the intrinsic set

Perou et al highlight of four gene clusters obtained by clustering their intrinsic set of 496 genes. Of these, cluster d is almost identical to our cluster G21, found by G1(S) (without DG, see website

http://www.weizmann.ac.il/home/cskela/perou/perou_all_samples2/rows/r1_ctwc_all_samples2_c1.html). Their cluster c has many genes, including the estrogen receptor and the gata-binding protein 3, in common with our cluster G4, also found by G1(S). The other two clusters, e and f, were not found in the G1(S1) procedure as stable clusters.

5.5.2 New results

My next goal was to use the power of CTWC in order to find new classifications, not observed by Perou et al. In order to select the interesting partitions, I sorted all the gene and sample clusters that were generated, according to their stability; each cluster is scored by a stability parameter; the higher it is, more significant (statistically) is the cluster. The analysis described until now consisted of clustering 'pairs' (O_i, F_i), where the objects (O_i) were the 65 tumors, denoted by S1, and the features (F_i) were different subsets (i.e. clusters) of genes GI, that were generated by clustering the entire set of genes (G1) based on S (the 84 samples) and on S1. The clustering analysis presented now differs, in that I are looking also at 'pairs' in which the objects are subsets of S1, denoted by SX (X=2, 3…7); the SX clusters are presented in Fig. 5.7. This way I am zooming in to deeper levels of the data.

5.5.2.1 Improved ER+/ER- classifiers

The BLUE cluster of Perou et al contains 36 tumors, 32 of which are ER+. Hence, when viewed as an ER+ classifier, the BLUE cluster has purity P=32/36=0.89.

Since the total number of ER+ tumors is 48, its efficiency as a classifier is E=32/48=0.66.When I clustered the tumors using the G4 genes, I got purity P=0.87 and E=0.75.

Using the genes of G30 (which are correlated with cell proliferation) I found a cluster with P=0.86 and E=0.79 with respect to ER+. I saw that most of the ER+ tumors have high expression of the G4 genes (as also noted by Perou et al) and intermediate expression of the G30 genes. ER- tumors have, in general, higher proliferation than most ER+. I did find a small group of tumors that were ER+ and with high proliferation (high G30) - these were characterized as ErbB2-.

5.5.2.2 Subpartitions of the ER+ samples

S2(G21) uses the ErbB2-related genes of cluster G21, to cluster S2 - a cluster of 22 tumors, 21 of which are ER+. I find a cluster of samples with low, and another with high expression levels of G21. As follow up to this finding, I used G21 to cluster all 48 ER+ samples, and found again separation into low and high expression levels

5.5.2.3 Correlating high proliferation with genetic/clinical markers

High proliferation (i.e. high expression of G46) is observed in three tumor clusters, that were characterized as ER-/ErbB2- (highest) and ER-/ErbB2+ and ER+/ErbB2-. That is, missing at least one of these two proteins may be correlated with high levels of growth.

5.5.2.4 Predicting response to doxorubicin treatment

The procedure **S3(G46)** yielded very clear sub-clusters. 25 out of the 29 tumors of S3 are ER+ and the remaining 4 are ER-. Clustering them using G46 yields 3 clear subgroups of low, intermediate and high expression levels of G46. The intermediate group contained all three "BEFORE" samples that responded well to doxorubicin treatment: the corresponding "AFTER" tumors were in the healthy-like, low expression cluster. Since this finding indicated that the G46 genes may serve as predictors of success or failure of doxorubicin treatment, I looked at the operation **S1(G46)**, i.e. clustered all samples using G46.

The **S1(G46)** analysis identified a cluster of 26 tumors with intermediate expression levels. Among these 10 were "BEFORE" (out of 20 such tumors). However, ALL 3 "BEFORE" tumors that responded positively to the treatments were in this cluster. The three matching "AFTER" samples were in another cluster of low proliferation, joining the "NORMAL" samples. Hence intermediate expression of the G46 genes may serve as a marker for a relatively high success rate of the doxorubicin treatment (3/10 versus 3/20 for the entire set of BEFORE samples).

5.5.2.5 ErbB2- Detector

The operation **S1(G9)**, which identified the homologue of the GREEN cluster, also yielded another cluster of samples, which had low expression of the G9 genes and contained most of the ErbB2- tumors (P=0.7, E=0.85).

5.5.2.6 Other findings

Interesting partitions were found by **S1(G13)**.

Now I represent the results in detail.

5.5.2.1 Improved ER+/ER- classifications

The BLUE cluster of Perou et al contains 36 tumors, 32 of which are ER+. Hence when viewed as an ER+ classifier, the BLUE cluster has purity P=0.89. Of the 48 ER+ tumors the BLUE cluster contains 32; hence its efficiency as a classifier is E=0.66.

When I cluster all the 65 tumors using the G4 genes, I get a cluster (cluster **a** of Fig. 5.10A) with purity 0.87 and efficiency 0.75 (see Table 5.5).

S1 (G4)	ER+ (48)	ER- (12)
Cluster a (41)	36 P=0.87 E=0.75	1
Cluster b (15)	9	4
Cluster c (9)	2	7
BLUE cluster of Perou et al (36)	32 P=0.89 E=0.66	2

Table 5.5. Distribution of the ER+/ER- tumors among the clusters generated by the S1(G4) procedure. Cluster a identifies ER+ with high purity and efficiency. The BLUE cluster is presented for comparison.

Using the genes of G30 (correlated with cell proliferation), I found a cluster with P=0.86 and E=0.79 (cluster c of Fig. 5.11A) with respect to ER+ (see Table 5.6). This suggests that most ER+ tumors may show low aggressiveness.

S1 (G30)	ER+ (48)	ER- (12)
Cluster a1 (11)	2	9
Cluster b (5)	5	0
Cluster c (44)	**38** P=0.86 E=0.79	3

Table 5.6. Distribution of the ER+/ER- tumors among the clusters generated by the S1(G30) procedure. Cluster c identifies ER+ with high purity and efficiency.

I see that the most common ER+ tumors have high expression of G4 genes (as also noted by Perou et al) and intermediate expression of G30 genes. ER- tumors have higher proliferation rate than most ER+. I did find a small group of tumors (part of cluster b – see Fig.11) that were ER+ and show high expression levels of G30 genes – these were characterized also as ErbB2-.

5.5.2.2 Subpartitions of the ER+ samples

Clustering S2 (22 samples) using G21; **S2(G21)**. The samples of S2 constitute a stable cluster, that was obtained by clustering all the 65 samples on the basis of the expression levels of all the genes, **S1(G1)**. All but one of the 22 tumors of S2, are ER+. When these 22 samples are clustered, based on G21 (ErbB2 cluster), S2 splits to two very stable clusters shown in Fig. 5.15A; in the samples of cluster **a** the genes of G21 have low expression levels whereas in cluster **b** their expression levels are high. The clusters in the dendrogram of Fig. 5.15A are colored according to their proportion of the ErbB2+ group (tumors that over expressed the ErbB2 protein, according to the supplementary information reported in Perou et al). Hence, the dendrogram of Fig.15A presents a new classification of the ER+ tumors; the 'Low Erb-B2 group' (cluster a), characterized by low expression level of the G21 genes, and the 'High Erb-B2 group' (cluster b), characterized by high expression level of the G21 genes. One can conclude, that one of the main reasons for the development of cancer in the tumors of the 'ER+/High Erb-B2' group is the over expression of the ErbB2 protein, mostly as a results of amplification of the locus that contains the corresponding ErbB2 gene. Therefore, a natural

question to ask is: what factor, or factors, may be involved with tumors that belong to the 'ER+/Low ErbB2' group?

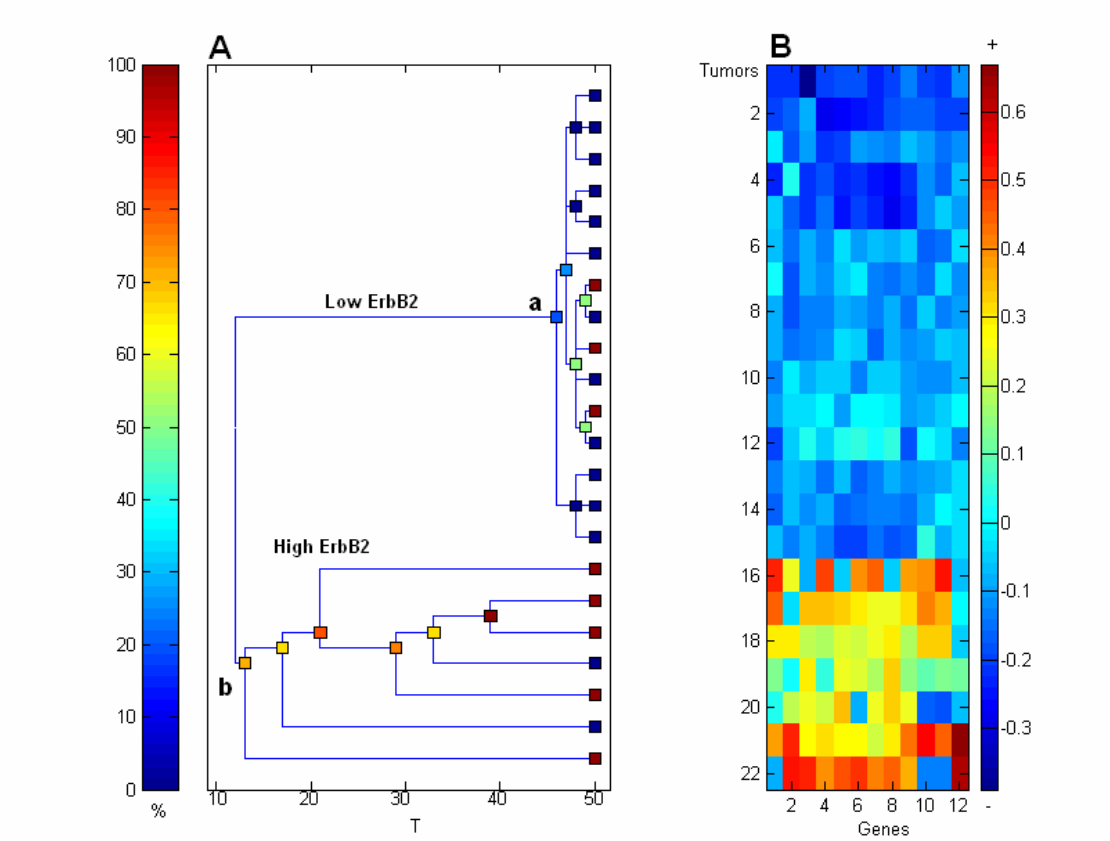

Figure 5.15. The S2(G21) procedure produces a very clear cluster with low expression levels. Boxes are colored according to the percentage of ErbB2+ tumors in the corresponding clusters. S2 is a sample cluster rich in ER+ tumors. All but one tumors in this group are ER+.

Clustering all 48 ER+ using G21; **ER+(G21)**. Next, I tested the power of G21 as an ER+ subdivision classifier, clustering all the 48 ER+ tumors using the G21 genes (ErbB2 cluster). Again, the ER+ tumors split into two clusters, shown in Fig. 5.16A; the samples of sub-cluster **a** have low expression levels whereas in cluster b their expression levels are high. The clusters in the dendrogram of Fig. 5.15A are colored according to their proportion of the ErbB2-group (tumors that lack the Erb-B2 protein). As one sees, the sub partitions of ER+ shown in Fig. 5.16A are similar to those obtained in Fig. 5.15A; in the bigger subgroup of ER+ tumors, contained in clusters **a** of Figs. 5.15A and 16A, the genes of G21 have low expression levels, and in the smaller subgroups, contained in clusters **b** of Figs. 5.15A and 16A, their expression levels are high.

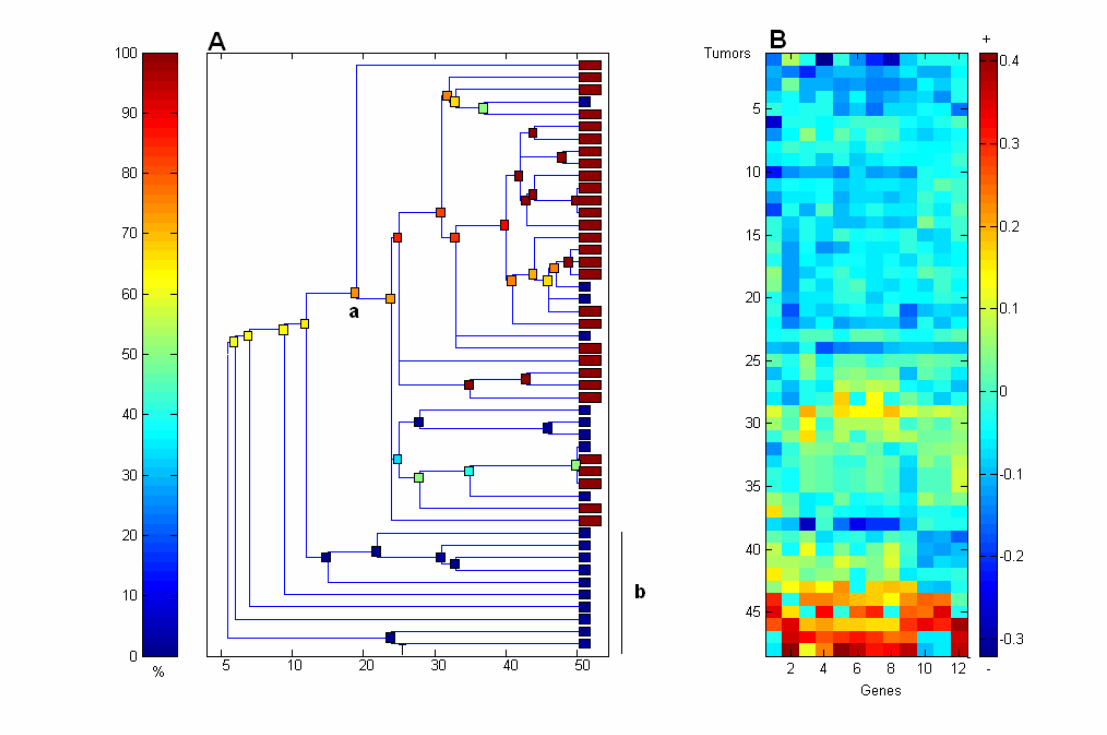

Figure 5.16. Clustering all ER+ tumors using the genes of G21 produces a large cluster a of tumors with low expression levels and a small cluster b with high expression. Here boxes are colored according to the percentage of Erb-B2- tumors in the corresponding clusters.

5.5.2.3 Correlating high proliferation with genetic/clinical markers

Genes correlated with proliferation (i.e. high expression of G46, see list in appendix) are observed in three tumor clusters of Fig. 5.12A (clusters **b**, **c** and **d**), that were characterized as ER-/ErbB2- (highest) and ER-/ErbB2+ and ER+/ErbB2-. That is, missing at least one of these two proteins correlate with aggressive growth.

Clustering S3 (29 sample) based on G46, **S3(G46)**, exhibit a sharper picture than obtained in Fig. 5.12A. In this operation, similarly to the features that were seen in Fig. 5.12A, most of the tumors that exhibit higher expression levels of G46 (clusters b and c of Fig. 5.16A), are either ER- or Erb-B2-, where the ER-/ErbB2- tumors (cluster c) exhibit the highest expression level of G46.

5.5.2.4 Predicting response to doxorubicin treatment, using the G46 set of genes

This example demonstrates the way I use CTWC. Our standard automated procedure (using stable gene clusters to cluster stable sample clusters) led us to look at the procedure S3(G46). The samples of S3 constitute a stable cluster, that was obtained by clustering all the 65 samples on the basis of the expression levels of all the genes, S1(G1). The S3 cluster contains the S2 cluster, described below. All but four of the 29 tumors of S3 are ER+, and the others are ER-. When these 29 samples are clustered, based on G46 (see Table 5.2), S3 splits into three very stable clusters shown in Fig. 5.17A, that exhibit low (**a**), middle (**b**) and high (**c**) expression levels of G46. The clusters in the dendrogram Fig. 5.17A are colored according to their proportion of the ER- group. Cluster **c** is the most aggressive, according to the proliferation rate of its cells. The red arrows if Fig. 5.16B point to the locations of the three 'responder' tumors before the doxorubicin treatment, and the black arrows point to the same tumors after the treatment. The 'responders' exhibited before the treatment intermediate expression levels of G46, and after the treatment their expression levels of G46 decreased, and they moved to the healthy-like, low expression cluster. This finding indicated that the G46 genes may serve as predictors of success or failure of doxorubicin treatment and hence I looked at the operation **S1(G46)**, i.e. clustered all samples using G46.

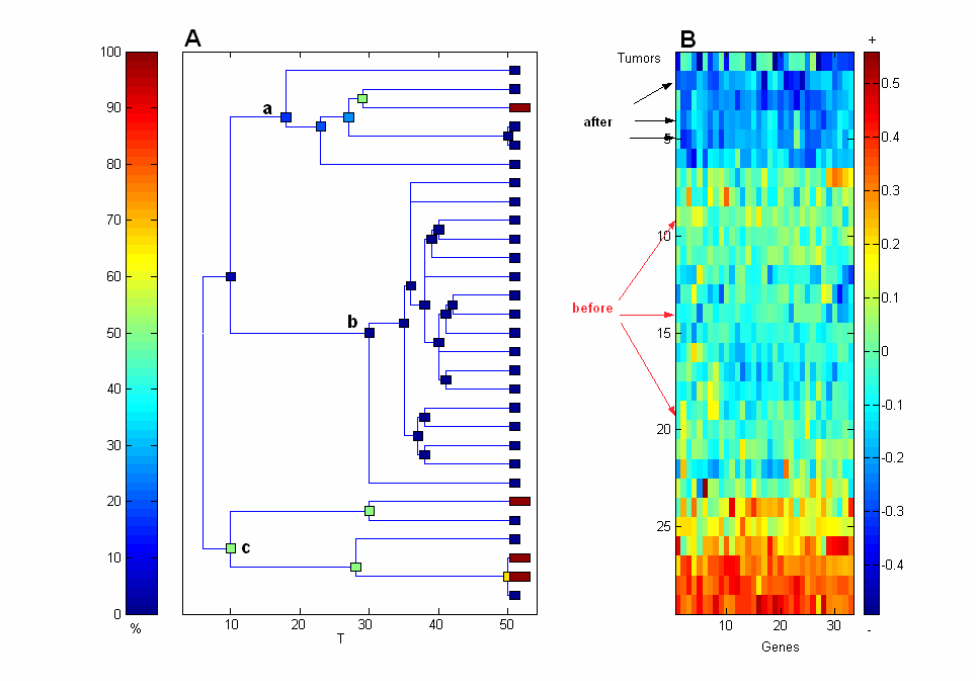

Figure 5.17. The S3(G46) clustering procedure yields three clearly separated clusters of high, intermediate and low expression levels. Boxes are colored according to their proportion of ER- tumors (red is ER). The arrows point to the tumors that responded to doxorubicin, red - before and black - after treatment.

The **S1(G46)** analysis (described above) identified a cluster of 26 tumors with intermediate expression levels (cluster b of Fig. 5.12A). Among these 10 were "BEFORE" (out of 20 such tumors). However, all the three "BEFORE" tumors that responded positively to the treatment were in this cluster. The three matching "AFTER" samples were in another cluster of low proliferation (cluster a of Fig. 5.12A), joining the "NORMAL" samples. Hence intermediate expression level of the G46 genes may serve as a marker for a relatively high success rate of the Doxorubicin treatment (3/10 versus 3/20 for the entire set of BEFORE samples).

5.5.2.5 ErbB2- detector

The operation **S1(G9)** (see Fig. 5.13), which identified the homologue of the GREEN cluster, also yielded another cluster of samples (cluster a of Fig. 5.13A), which had low expression of the G9 genes and contained most of the ErbB2- tumors (P=0.7, E=0.85).

5.5.2.6 Other findings

Clustering S1 based on G13; **S1(G13)**. When the 65 samples (S1) are clustered, based on G13 (see Table 5.2 and appendix), S1 splits to two main groups. One group of samples exhibits higher expression levels of G13; this group includes clusters **a**, and **b** (see Fig. 5.18A). The second group exhibits lower expression levels of G13 and includes the main cluster located in the center of the dendrogram. The high expression group contains the NORMAL samples and the 3 ‘responder’ tumors after the treatment. Before the treatment the 3 ‘responder’ tumors exhibit lower expression levels of G13 (pointed by 3 red arrows). The clusters in the dendrogram Fig. 5.18A are colored according to their content of the tumors containing low level of proliferation genes (cluster **a** of Fig. 5.12A); these samples show low expression levels of G46 (the ‘proliferation genes’). These samples fell also in clusters **a** and **b** of Fig. 5.18A, which may indicate an opposite relation between the expression levels of the cluster G46 to those of the cluster G13. G13 included 2 genes, JunB and Fos, which participate in controlling entrance to the cell cycle process. However it is known that Fos is activated by p53 and may have other functions related to suppression. These genes belong to two large genes families; Jun and Fos. These families are encoded by genes that respond to extra cellular primary growth factors, which lead the generation of Jun-Fos transcription factor complexes, known as AP-1. Different members of Jun and Fos can generate divers complexes of AP-1, which usually promote cell cycle proliferation.

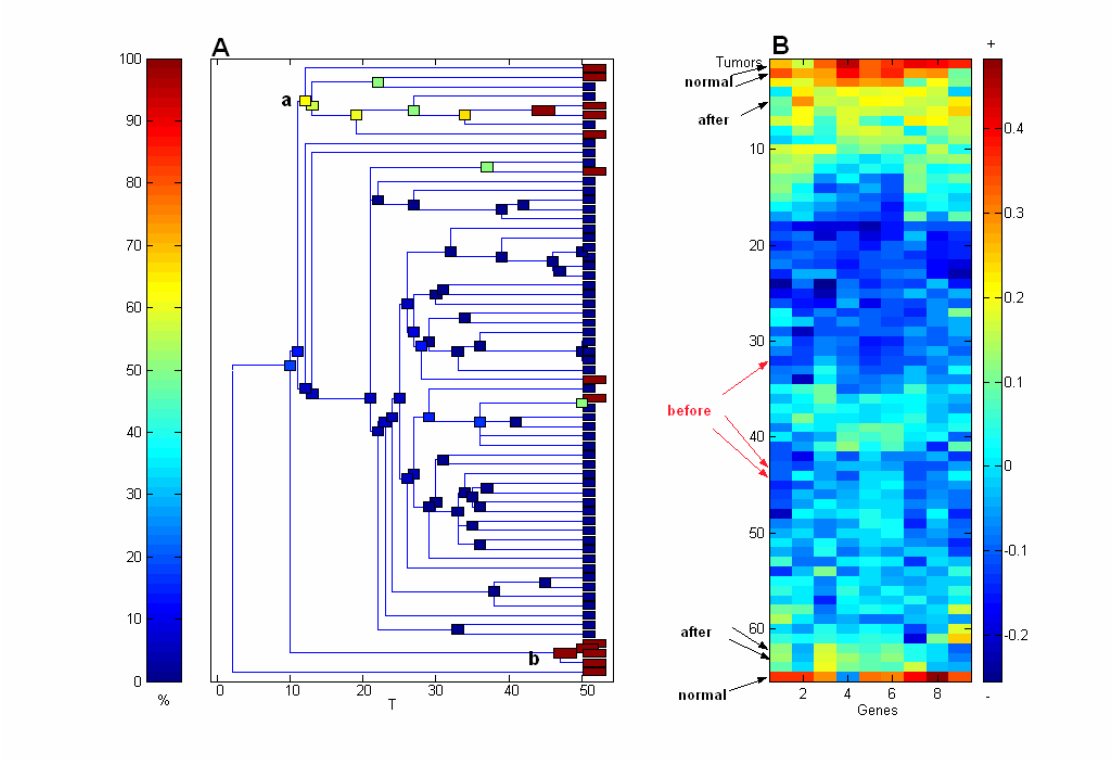

Figure 5.18. S1(G13) ; the majority of the tumors have low expression levels. Those of clusters a and b have higher expression of G13, but low expression of the genes of the proliferation cluster G46. The tumors labeled red are those that belong to cluster a of Fig 12A, i.e. those with low expression of the G46 genes.

Bibliography

[1] Ferguson, J. A., Boles, T. C., Adams, C. P. & Walt, D. R. (1996). Nature Biotechnol. 14, 1681-1684.

[2] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. (1998). Mol Biol Cell. Dec;9(12):3273-97.

[3] Liang P, Pardee AB. (1992). Science Aug 14;257(5072):967-71.

[4] Velculescu, V.E. Zhang, L., Vogelestein, B., and Kinzler, K.W. (1995). Science 270:484 -487.

[5] Ramsay. G. (1998). Nature Biotechnology, 16:40-44.

[6] Schema, M., Shalon, D., Davis, R.W., and brown, P.O. (1995). Sience 270, 467-470.

[7] Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. & Brown, E., (1996) Nat. Biotechnol. 14, 1675 -1680.

[8] AC Pease, D Solas, EJ Sullivan, MT Cronin, CP Holmes, and SPA Fodor. (1994).Proc. Natl. Acad. Sci. USA. 91, 5022-5026.

[9] Schulze A., Downward J. (2001). Nat Cell Biol. August, 3(8), 190-5.

[10] Lipshutz RJ., Fodor SP., Gingeras TR., Lockhart DJ. (1999). Nat Genet. anuary, 21(1 Suppl), 20-4.

[11] Lockhart D.J., Winzeler E.A. (2000). Nature, 405, 827-835. 15.

[12] Blatt M, Wiseman S and Domany E. (1996). Physical Rev. Lett.; 76: 3251-3254.

[13] Getz. G., Levine E, Domany E and Zhang MQ. (2000). Physica A.; 279: 57-464.

[14] Getz G., Levine E., Domany E. (2000). Proc Natl Acad Sci U S A.Oct 24;97(22):12079-84.

[15] Alberts B., Bray D., Lewis J., Raff M., Roberts K., Watson D. (1994). Biology of the ell, Third edition.

[16] El-Deiry WS, Kern SE, Pietenpol JA, Kinzler KW and Vogelstein B. (1992). Nat. Genet. , 45-49.

[17] Levine AJ. (1997). *Cell* 88, 323-331.

[18] El-Deiry WS, Kern SE, Pietenpol JA, Kinzler KW and Vogelstein B. (1992). Nat. Genet. , 45-49.

[19] Paul H. Kussie, Svetlana Gorina, Vincent Marechal, Brian Elenbaas, Jacque Moreau,

rnold J. Levine, Nikola P. Pavletich. (1996). Science November 8; 274: 948-953.

[20] Cho Y, Gorina S, Jeffrey PD, Pavletich NP**.** (1994). Science 1994 July 15; 265:346-355.

[21] Yu J, Zhang L, Hwang PM, Rago C, Kinzler KW and Vogelstein B. (1999). Proc. Natl. Acad. Sci. USA 96, 14517-14522.

[22] Michalovitz D, Halevy O and Oren M. (1990). Cell; 62: 671-680.

[23] Hilary A. Coller, Carla Grandori, Pablo Tamayo, Trent Colbert, Eric S. Lander, Robert N. Eisenman, and Todd R. (2000). Proc. Natl. Acad. Sci.; 97: 3260-3265.

[24] Rónán C. O'Hagan, Nicole Schreiber-Agus, Ken Chen, Gregory David, Jeffrey A. Engelman, Richard Schwab, Leila Alland, Cole Thomson, Donald R. Ronning, James C. Sacchettini, Paul Meltzer, Ronald A. DePinho. (2000). Nature Genetics; 24: 113-119.

[25] Kannan K, Amariglio N, Rechavi G, Jakob-Hirsch J, Kela I, Kaminski N, Getz G, Domany E, Givol D. (2001). Oncogene April; 20: 2225-2234.

[26] Ginsberg D, Michalovitz D, Ginsberg D and Oren, M. (1991). Mol. Cell Biol; 11: 582 585.

[27] Haupt Y, Maya R, Kazaz A and Oren, M. (1997). Nature; 387: 296-299.

[28] Kubbutat MH, Jones SN and Vousden KH. (1997). Nature 387, 299-303.

[29] Blatt M, Wiseman S and Domany E. (1996). Physical Rev. Lett.; 76: 3251-3254.

[30] Getz G, Levine E, Domany E and Zhang MQ. (2000). *Physica A. 2000;* 279: 457-464.

[31] Shivakumar CV, Brown DR, Deb S and Deb SP. (1995). Mol. Cell Biol; 15: 6785-6793.

[32] Zhan QM, Chen IT, Antinore MJ and Fornace AJ. (1998). Mol. Cell biol.; 18: 2768 2778.

[33] Sherlock G. (2000). Curr. Opin. Immunol.; 275: 37469-37473.

[34] Young RA. (2000). Cell 2000; 102: 9-15.

[35] Allred DC, Mohsin SK, Fuqua SA. (2001). Endocr Relat Cancer Mar;8(1):47-61

[36] Bodis S, Siziopikou KP, Schnitt SJ, Harris JR, Fisher DE. (1996). Cancer May 1;77(9):1831-5

[37] Harn HJ, Shen KL, Yueh KC, Ho LI, Yu JC, Chiu SC, Lee WH. (1997). Histopathology Dec;31(6):534-9

[38] Fuqua SA, Wiltschke C, Zhang QX, Borg A, Castles CG, Friedrichs WE, Hopp T, Hilsenbeck S, Mohsin S, O'Connell P, Allred DC. (2000). Cancer Res. Aug 1;60(15):4026-9

[39] Clarke RB, Howell A, Potten CS, Anderson E. (1997). Cancer Res. Nov 15;57(22): 4987

-91. (2001). J Mol Med. Oct;79(10):566-73.

[40] Pavelic K, Gall-Troselj K. (2001). J Mol Med Oct;79(10):566-73

[41] Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC, Lashkari D, Shalon D, Brown PO, Botstein D. (1999). Proc. Natl. Acad. Sci. U S A, 3;96(16):9212-7.

[42] Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR,Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. (2000). Nature 17;406(6797):747-52.

[43] Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL. (2001). Proc. Natl. Acad. Sci.11;98(19):10869-74.

[44] Ahr A, Holtrich U, Solbach C, Scharl A, Strebhardt K, Karn T, Kaufmann M. (2001). J Pathol. Oct;195(3):312-20.

[45] Seshadri R, Firgaira FA, Horsfall DJ, McCaul K, Setlur V, Kitchen P. (1993). J Clin Oncol. Oct;11(10):1936-42.

[46] Tandon AK, Clark GM, Chamness GC, Ullrich A, McGuire WL. (1989). J Clin Oncol. Aug;7(8):1120-8.

[47] West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR, Nevins JR. (2001). Proc. Natl. Acad. Sci. U S A 25; 98(20): 11462-7.

[48] Lonning PE, Sorlie T, Perou CM, Brown PO, Botstein D, Borresen-DaleAL. (2001) Endocr Relat Cancer. Sep;8(3):259-63.

[49] Geisler S, Lonning PE, Aas T, Johnsen H, Fluge O, Haugen DF, Lillehaug JR, Akslen LA, Borresen-Dale AL. (2001). Cancer Res. Mar 15;61(6): 2505-12.

[50] Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, SchwabM, Antonescu CR, Peterson C, Meltzer PS.(2001). Nat. Med. Jun; 7(6): 673-9.

Appendix

Cluster G4

1	1593	X-BOX BINDING PROTEIN 1
2	1594	"HEPATOCYTE NUCLEAR FACTOR 3, ALPHA"
3	1595	GATA-BINDING PROTEIN 3
4	1596	GATA-BINDING PROTEIN 3
5	1597	GATA-BINDING PROTEIN 3
6	1598	GATA-BINDING PROTEIN 3
7	1599	ESTROGEN RECEPTOR 1
8	1600	ESTROGEN RECEPTOR 1
9	1601	ANNEXIN XXXI
10	1602	ANNEXIN XXXI

Cluster G9

1	647	"ESTS, WEAKLY SIMILAR TO W01A11.2 GENE PRODUCT [C.ELEGANS]"
2	658	78946
3	659	"FATTY ACID BINDING PROTEIN 4, ADIPOCYTE"
4	660	"FATTY ACID BINDING PROTEIN 4, ADIPOCYTE"
5	665	GLUTATHIONE PEROXIDASE 3 (PLASMA)
6	667	"ALCOHOL DEHYDROGENASE 2 (CLASS I), BETA POLYPEPTIDE"
7	668	AQUAPORIN 7
8	669	484535
9	670	LIPOPROTEIN LIPASE
10	671	GLYCEROL-3-PHOSPHATE DEHYDROGENASE 1 (SOLUBLE)
11	672	"RETINOL-BINDING PROTEIN 4, INTERSTITIAL"
12	673	"INTEGRIN, ALPHA 7"
13	674	85660

Cluster G11

1	55	"HOMO SAPIENS MRNA FOR KIAA0758 PROTEIN, PARTIAL CDS"
2	62	"TEK TYROSINE KINASE, ENDOTHELIAL (VENOUS MALFORMATIONS, MULTIPLE CUTANEOUS AND MUCOSAL)"
3	63	LIM BINDING DOMAIN 2
4	64	KINASE SCAFFOLD PROTEIN GRAVIN
5	65	359722
6	66	TYROSINE KINASE WITH IMMUNOGLOBULIN AND EPIDERMAL GROWTH FACTOR HOMOLOGY DOMAINS
7	67	CD34 ANTIGEN
8	68	"HUMAN DNA SEQUENCE FROM CLONE 1033B10 ON CHROMOSOME 6P21.2-21.31. CONTAINS THE BING5 GENE, EXONS 11 TO 15 OF THE BING4 GENE, THE GENE FOR GALT3 (BETA3-GALACTOSYLTRANSFERASE), THE RPS18 (40S RIBOSOMAL PROTEIN S18) GENE, THE SACM2"
9	69	69672
10	70	"HOMO SAPIENS KDR/FLK-1 PROTEIN MRNA, COMPLETE CDS"
11	71	"LAMININ, ALPHA 4"

Cluster G12

1	425	"LAMININ, GAMMA 2 (NICEIN (100KD), KALININ (105KD), BM600 (100KD), HERLITZ JUNCTIONAL EPIDERMOLYSIS BULLOSA))"
2	426	ANNEXIN VIII
3	427	"ESTS, HIGHLY SIMILAR TO PROBABLE ATAXIA-TELANGIECTASIA GROUP D PROTEIN [H.SAPIENS]"
4	428	KERATIN 17
5	429	KERATIN 17
6	430	"ESTS, HIGHLY SIMILAR TO KERATIN K5, 58K TYPE II, EPIDERMAL [H.SAPIENS]"
7	431	KERATIN 5 (EPIDERMOLYSIS BULLOSA SIMPLEX DOWLING-MEARA/KOBNER/WEBER-COCKAYNE TYPES)
8	432	"ESTS, HIGHLY SIMILAR TO KERATIN K5, 58K TYPE II, EPIDERMAL"
9	433	BULLOUS PEMPHIGOID ANTIGEN 1 (230/240KD)
10	434	S100 CALCIUM-BINDING PROTEIN A2
11	435	"INTEGRIN, BETA 4"
12	436	"INTEGRIN, BETA 4"
13	437	2255577
14	438	"LAMININ, ALPHA 3 (NICEIN (150KD), KALININ (165KD), BM600 (150KD), EPILEGRIN)"
15	439	"COLLAGEN, TYPE XVII, ALPHA 1"
16	440	BASONUCLIN

Cluster G15

1	647	"ESTS, WEAKLY SIMILAR TO W01A11.2 GENE PRODUCT [C.ELEGANS]"
2	658	78946
3	659	"FATTY ACID BINDING PROTEIN 4, ADIPOCYTE"
4	660	"FATTY ACID BINDING PROTEIN 4, ADIPOCYTE"
5	661	"FATTY ACID BINDING PROTEIN 4, ADIPOCYTE"
6	662	"MDGI/FATTY ACID BINDING PROTEIN 3, MUSCLE AND HEART"
7	663	"CD36 ANTIGEN (COLLAGEN TYPE I RECEPTOR, THROMBOSPONDIN RECEPTOR)"
8	665	GLUTATHIONE PEROXIDASE 3 (PLASMA)
9	667	"ALCOHOL DEHYDROGENASE 2 (CLASS I), BETA POLYPEPTIDE"
10	668	AQUAPORIN 7
11	669	484535
12	670	LIPOPROTEIN LIPASE
13	671	GLYCEROL-3-PHOSPHATE DEHYDROGENASE 1 (SOLUBLE)
14	672	"RETINOL-BINDING PROTEIN 4, INTERSTITIAL"
15	673	"INTEGRIN, ALPHA 7"
16	674	85660

Cluster G16

1	178	471748
2	180	TRANSGELIN/SM22
3	181	SMOOTH MUSCLE PROTEIN 22-ALPHA
4	182	LUMICAN
5	183	FIBULIN 1
6	184	"COLLAGEN, TYPE VI, ALPHA 3"
7	185	"HOMO SAPIENS OSF-2 MRNA FOR OSTEOBLAST SPECIFIC FACTOR 2 (OSF-2P1), COMPLETE CDS"
8	186	"COLLAGEN, TYPE III, ALPHA 1 (EHLERS-DANLOS SYNDROME TYPE IV, AUTOSOMAL DOMINANT)"
9	187	"COLLAGEN, TYPE I, ALPHA 1"
10	188	"COLLAGEN, TYPE I, ALPHA 2"
11	189	"COLLAGEN, TYPE III, ALPHA 1 (EHLERS-DANLOS SYNDROME TYPE IV, AUTOSOMAL DOMINANT)"
12	190	"COLLAGEN, TYPE III, ALPHA 1 (EHLERS-DANLOS SYNDROME TYPE IV, AUTOSOMAL DOMINANT)"
13	191	"COLLAGEN, TYPE I, ALPHA 2"
14	192	THY-1 CELL SURFACE ANTIGEN
15	193	"HOMO SAPIENS, ALPHA-1 (VI) COLLAGEN"
16	194	"COLLAGEN, TYPE VI, ALPHA 1"
17	195	"COLLAGEN, TYPE VI, ALPHA 1"
18	196	"HUMAN ALPHA-2 COLLAGEN TYPE VI MRNA, 3' END"

Cluster G17

1	826	126449
2	832	KIAA0101 GENE PRODUCT
3	840	HOMO SAPIENS MRNA; CDNA DKFZP434F222 (FROM CLONE DKFZP434F222)
4	841	MINICHROMOSOME MAINTENANCE DEFICIENT (S. CEREVISIAE) 3
5	842	KIAA0166 GENE PRODUCT
6	843	NUCLEAR AUTOANTIGENIC SPERM PROTEIN (HISTONE-BINDING)
7	844	DEAD/H (ASP-GLU-ALA-ASP/HIS) BOX POLYPEPTIDE 11 (S.CEREVISIAE CHL1-LIKE HELICASE)
8	845	DEAD/H (ASP-GLU-ALA-ASP/HIS) BOX POLYPEPTIDE 11 (S.CEREVISIAE CHL1-LIKE HELICASE)
9	846	"MINICHROMOSOME MAINTENANCE DEFICIENT (MIS5, S. POMBE) 6"
10	847	DNA (CYTOSINE-5-)-METHYLTRANSFERASE 1
11	848	"ESTS, MODERATELY SIMILAR TO !!!! ALU SUBFAMILY J WARNING ENTRY !!!! [H.SAPIENS]"
12	849	ENHANCER OF ZESTE (DROSOPHILA) HOMOLOG 2
13	850	REPLICATION FACTOR C (ACTIVATOR 1) 4 (37KD)
14	852	RIBONUCLEOTIDE REDUCTASE M2 POLYPEPTIDE
15	853	CDC28 PROTEIN KINASE 2
16	854	CDC28 PROTEIN KINASE 2
17	855	PITUITARY TUMOR-TRANSFORMING 1
18	856	"HUMAN UBIQUITIN CARRIER PROTEIN (E2-EPF) MRNA, COMPLETE CDS"
19	857	"ESTS, HIGHLY SIMILAR TO MITOTIC KINESIN-LIKE PROTEIN-1 [H.SAPIENS]"
20	858	TROPHININ-ASSISTING PROTEIN (TASTIN)
21	859	"CELL DIVISION CYCLE 20, S.CEREVISIAE HOMOLOG"
22	860	"HOMO SAPIENS MRNA FOR KIAA0788 PROTEIN, PARTIAL CDS"
23	861	"HOMO SAPIENS HPV16 E1 PROTEIN BINDING PROTEIN MRNA, COMPLETE CDS"
24	862	"CENTROMERE PROTEIN F (350/400KD, MITOSIN)"
25	863	782283
26	864	POLYMYOSITIS/SCLERODERMA AUTOANTIGEN 1 (75KD)
27	865	CYCLIN A2
28	866	"HOMO SAPIENS MRNA FOR CDC2 DELTA T, COMPLETE CDS"
29	867	PROTEIN KINASE MITOGEN- ACTIVATED 13
30	868	POLO (DROSOPHIA)-LIKE KINASE
31	869	"HUMAN MRNA FOR KIAA0074 GENE, PARTIAL CDS"
32	870	BUDDING UNINHIBITED BY BENZIMIDAZOLES 1 (YEAST HOMOLOG)
33	871	MINICHROMOSOME MAINTENANCE DEFICIENT (S. CEREVISIAE) 4
34	872	FLAP STRUCTURE-SPECIFIC ENDONUCLEASE 1
35	873	236142
36	874	FORKHEAD (DROSOPHILA)-LIKE 16
37	876	SMALL NUCLEAR RIBONUCLEOPROTEIN POLYPEPTIDES B AND B1
38	877	PROLIFERATING CELL NUCLEAR ANTIGEN
39	878	PROLIFERATING CELL NUCLEAR ANTIGEN
40	880	"NON-METASTATIC CELLS 1, PROTEIN (NM23A) EXPRESSED IN"
41	881	"HUMAN MRNA FOR KIAA0098 GENE, PARTIAL CDS"
42	886	MEMBRANE-ASSOCIATED TYROSINE- AND THREONINE-SPECIFIC CDC2-INHIBITORY KINASE
43	889	244205

Cluster 21

1	1709	418240
2	1710	KIAA0130 GENE PRODUCT
3	1711	ERBB-2 RECEPTOR PROTEIN-TYROSINE KINASE PRECURSOR
4	1712	STEROIDOGENIC ACUTE REGULATORY PROTEIN RELATED
5	1713	ERBB2-POLYA
6	1714	V-ERB-B2 AVIAN ERYTHROBLASTIC LEUKEMIA VIRAL ONCOGENE HOMOLOG 2 (NEURO/GLIOBLASTOMA DERIVED ONCOGENE HOMOLOG)
7	1715	V-ERB-B2 AVIAN ERYTHROBLASTIC LEUKEMIA VIRAL ONCOGENE HOMOLOG 2 (NEURO/GLIOBLASTOMA DERIVED ONCOGENE HOMOLOG)
8	1716	ERBB2
9	1717	GROWTH FACTOR RECEPTOR-BOUND PROTEIN 7
10	1718	68400
11	1719	68400
12	1720	"SWI/SNF RELATED, MATRIX ASSOCIATED, ACTIN DEPENDENT REGULATOR OF CHROMATIN, SUBFAMILY E, MEMBER 1"

Cluster G23

1	600	"CD8 ANTIGEN, BETA POLYPEPTIDE 1 (P37)"
2	601	V-MAF MUSCULOAPONEUROTIC FIBROSARCOMA (AVIAN) ONCOGENE HOMOLOG
3	602	CHLORIDE INTRACELLULAR CHANNEL 2
4	605	IMMUNOGLOBULIN GAMMA 3 (GM MARKER)
5	606	COLONY STIMULATING FACTOR 1 (MACROPHAGE)
6	607	"NEUTROPHIL CYTOSOLIC FACTOR 1 (47KD, CHRONIC GRANULOMATOUS DISEASE, AUTOSOMAL 1)"
7	608	IMMUNOGLOBULIN LAMBDA-LIKE POLYPEPTIDE 2
8	609	IMMUNOGLOBULIN LAMBDA LIGHT CHAIN
9	610	HUMAN REARRANGED IMMUNOGLOBULIN LAMBDA LIGHT CHAIN MRNA
10	611	HUMAN REARRANGED IMMUNOGLOBULIN LAMBDA LIGHT CHAIN MRNA
11	612	HUMAN IG J CHAIN GENE
12	613	IMMUNOGLOBULIN J CHAIN
13	614	HUMAN IG J CHAIN GENE
14	615	"MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS II, DQ BETA 1"
15	616	IMMUNOGLOBULIN MU
16	618	MAX-INTERACTING PROTEIN 1
17	1080	CHEMOKINE (C-C MOTIF) RECEPTOR 2

Cluster G30

1	856	"HUMAN UBIQUITIN CARRIER PROTEIN (E2-EPF) MRNA, COMPLETE CDS"
2	857	"ESTS, HIGHLY SIMILAR TO MITOTIC KINESIN-LIKE PROTEIN-1 [H.SAPIENS]"
3	859	"CELL DIVISION CYCLE 20, S.CEREVISIAE HOMOLOG"
4	860	"HOMO SAPIENS MRNA FOR KIAA0788 PROTEIN, PARTIAL CDS"
5	861	"HOMO SAPIENS HPV16 E1 PROTEIN BINDING PROTEIN MRNA, COMPLETE CDS"
6	862	"CENTROMERE PROTEIN F (350/400KD, MITOSIN)"
7	863	782283
8	865	CYCLIN A2
9	866	"HOMO SAPIENS MRNA FOR CDC2 DELTA T, COMPLETE CDS"
10	867	PROTEIN KINASE MITOGEN- ACTIVATED 13
11	868	POLO (DROSOPHIA)-LIKE KINASE
12	869	"HUMAN MRNA FOR KIAA0074 GENE, PARTIAL CDS"
13	870	BUDDING UNINHIBITED BY BENZIMIDAZOLES 1 (YEAST HOMOLOG)
14	871	MINICHROMOSOME MAINTENANCE DEFICIENT (S. CEREVISIAE) 4
15	874	FORKHEAD (DROSOPHILA)-LIKE 16

Cluster G35

1	605	IMMUNOGLOBULIN GAMMA 3 (GM MARKER)
2	607	"NEUTROPHIL CYTOSOLIC FACTOR 1 (47KD, CHRONIC GRANULOMATOUS DISEASE, AUTOSOMAL 1)"
3	608	IMMUNOGLOBULIN LAMBDA-LIKE POLYPEPTIDE 2
4	609	IMMUNOGLOBULIN LAMBDA LIGHT CHAIN
5	610	HUMAN REARRANGED IMMUNOGLOBULIN LAMBDA LIGHT CHAIN MRNA
6	611	HUMAN REARRANGED IMMUNOGLOBULIN LAMBDA LIGHT CHAIN MRNA
7	614	HUMAN IG J CHAIN GENE
8	616	IMMUNOGLOBULIN MU
9	617	EARLY DEVELOPMENT REGULATOR 2 (HOMOLOG OF POLYHOMEOTIC 2)
10	618	MAX-INTERACTING PROTEIN 1
11	1089	CD79A ANTIGEN (IMMUNOGLOBULIN-ASSOCIATED ALPHA)

Cluster G46

1	827	APOPTOSIS INHIBITOR 4 (SURVIVIN)
2	830	258761
3	832	KIAA0101 GENE PRODUCT
4	846	"MINICHROMOSOME MAINTENANCE DEFICIENT (MIS5, S. POMBE) 6"
5	850	REPLICATION FACTOR C (ACTIVATOR 1) 4 (37KD)
6	851	V-MYB AVIAN MYELOBLASTOSIS VIRAL ONCOGENE HOMOLOG-LIKE 2
7	853	CDC28 PROTEIN KINASE 2
8	854	CDC28 PROTEIN KINASE 2
9	855	PITUITARY TUMOR-TRANSFORMING 1
10	856	"HUMAN UBIQUITIN CARRIER PROTEIN (E2-EPF) MRNA, COMPLETE CDS"
11	857	"ESTS, HIGHLY SIMILAR TO MITOTIC KINESIN-LIKE PROTEIN-1 [H.SAPIENS]"
12	858	TROPHININ-ASSISTING PROTEIN (TASTIN)
13	859	"CELL DIVISION CYCLE 20, S.CEREVISIAE HOMOLOG"
14	860	"HOMO SAPIENS MRNA FOR KIAA0788 PROTEIN, PARTIAL CDS"
15	861	"HOMO SAPIENS HPV16 E1 PROTEIN BINDING PROTEIN MRNA, COMPLETE CDS"
16	862	"CENTROMERE PROTEIN F (350/400KD, MITOSIN)"
17	863	782283
18	864	POLYMYOSITIS/SCLERODERMA AUTOANTIGEN 1 (75KD)
19	865	CYCLIN A2
20	866	"HOMO SAPIENS MRNA FOR CDC2 DELTA T, COMPLETE CDS"
21	867	PROTEIN KINASE MITOGEN- ACTIVATED 13
22	868	POLO (DROSOPHIA)-LIKE KINASE
23	869	"HUMAN MRNA FOR KIAA0074 GENE, PARTIAL CDS"
24	870	BUDDING UNINHIBITED BY BENZIMIDAZOLES 1 (YEAST HOMOLOG)
25	871	MINICHROMOSOME MAINTENANCE DEFICIENT (S. CEREVISIAE) 4
26	872	FLAP STRUCTURE-SPECIFIC ENDONUCLEASE 1
27	873	236142
28	874	FORKHEAD (DROSOPHILA)-LIKE 16
29	876	SMALL NUCLEAR RIBONUCLEOPROTEIN POLYPEPTIDES B AND B1
30	877	PROLIFERATING CELL NUCLEAR ANTIGEN
31	878	PROLIFERATING CELL NUCLEAR ANTIGEN
32	880	"NON-METASTATIC CELLS 1, PROTEIN (NM23A) EXPRESSED IN"
33	881	"HUMAN MRNA FOR KIAA0098 GENE, PARTIAL CDS"