Systems Biology: Gene Expression Signatures and the Regulation of Transcription

Ph.D. theses

Dr. Győrffy András

Semmelweis University
Clinical Medicine Doctoral School

Tutors: Dr. Tulassay Zsolt D.Sc.
Dr. Győrffy Balázs, Ph.D.

Opponents: Dr. Prohászka Zoltán, PhD.
Dr. Treszl András, Ph.D.

Board of final examination: Dr. Rácz Károly med.habil, D.Sc.
Members of board: Dr. Nagy Gábor, Ph.D.
Dr. Hermann Róbert, Ph.D.
Dr. Banai János med.habil.

Budapest
2007

Title: Systems Biology: Gene Expression Signatures and the Regulation of Transcription
Author: Dr. Győrffy András
Tutors: Dr. Tulassay Zsolt and Dr. Győrffy Balázs SE II. Department of Internal Medicine, 2007

1. INTRODUCTION

Due to the spread of microarray technologies we have entered into possession of a vast knowledge about the degree of transcription occurring in the normal and diseased cells and tissues. The resulting databanks can be mined by appropriate statistical methods to gain exact information about the relative expression changes of the specific genes. However our knowledge regarding the diversified common controlling networks for all of these genes is scarce.

Studying the regulating mechanisms of these expression changes of genes made up the purpose of my Ph. D. work within the confines of which I was engaged in two alternative *in silico* research projects. During this research program I investigated both *cis* and *trans* transcription regulations. The transcription factors that are responsible for influencing the transcription belong to the examined *trans* regulatory elements. I investigated their effect via gene lists that have been correlated with doxorubicin and 5-fluorouracil chemotherapy resistance during earlier research projects. Based on my research the E47 transcription factor bears an important role in the co-regulation of gene expression signatures associated with doxorubicin resistance in breast cancer. Discovering the modulation of the apoptic machinery was identified as a key mechanism in the formation of 5-fluorouracil resistance.

The manner of *cis* regulation was investigated focusing on the analysis of antisense transcription. The antisense transcription was measured by rating the gene expression data of 1182 mice. Based on the obtained data I estimated the antisense transcription to be of 43% full genome-wide frequency. Compared to the other chromosomes I found a relatively high antisense expression on the chromosomes 14 and 1. At high expression levels an inverse sense-antisense gene expression correlation was observed. These results support the regulatory characteristics of antisense transcription in mouse genomes.

During these studies, other than the research tasks, I also performed such a technological development that enables the adaptation of the diagnostical measurements of the obtained and simplified microarray results. A positioning system and an automated pipetting system were developed. The latter applies disposable pipette tips commercially available. Although this development and creating the related sample protection and patent took a significant part of the effective work, I summarized its results only briefly in the ’Methods’ and ’Results’ chapters. The reason for this is that irrespectively of the results they bear no relevance in the evaluation of the Ph. D. thesis.

2. OBJECTIVES

1. Establishing a consensus gene list based on microarray data

My primary objective was to set up a consensus gene list where the causal role of specific genes was verified during the full genome-wide experiments. This was based on gene expression signatures correlated with doxorubicin resistancy experienced during tumor chemotherapy.

2. Identifying transcription factors

We can assume that the co-regulated consensus gene list is backed up by common regulatory mechanisms. The goal was determined as to identify the common transcription factor binding areas in correlation with doxorubicin resistancy and to set up the resistancy systems biology model with their help.

3. Antisense transcription in gene expression signatures

In my study the objective I proposed was the microarray based investigation of the expressions of the reverse complementary sequencies in order to examine the sense and antisense transcription on a more comprehensive genomic level. The purpose was to describe the frequency of the antisense transcription and the possible regulation characteristics based on the obtained data. A further aim of my disquisition was the research of the potential specific distribution signatures of the regulatory elements, in the interest of which I investigated the chromosome-specific expression signatures of the sense and antisense transcription.

4. Technological development

The applied microarray technologies cannot be used on a routine basis due to their high price and complexity. My purpose was the evolvement of an automated measuring system i.e. an automated pipette equipment that would make a cheap and simple pipetting on an arbitrary surface possible and these surfaces would also serve as places for performing electro-chemical and molecular measurements. My goal was the creation of a positioning unit that would define the relative and absolute position of objects. Determining the position of the pipette tips played an emphasized role during the development research.

3. METHODS

1. Scheme of source data for searching transcription factors

I was looking for publications in the PubMed databank that focus on doxorubicin resistancy. For search I used the English words ’doxorubicin’, ’cancer’, ’gene expression’ and ’microarray’. I excluded publications that were restricted to specific cell lines from the search since only given cell-line-specific mechanisms were shown in these, being narrowed down in the resistancy models. In order to decrease heterogenity I also excluded the publications that examined responses to several treatments in clinical samples and the ones that were not directly connected to the substance of the research project, like reviews for example. Reckoning with all this I took three different analyses related to doxorubicin resistancy into consideration, together with the gene lists and microarray platforms belonging to them. In these publications Affymetrix and Stanford chips were used. I connected the genes from the two platforms based on the Genebank Accession Number, that I downloaded from the Netaffx database. I selected the recurring genes with the help of Microsoft Access. Hereafter I dealt exclusively with recurring genes.

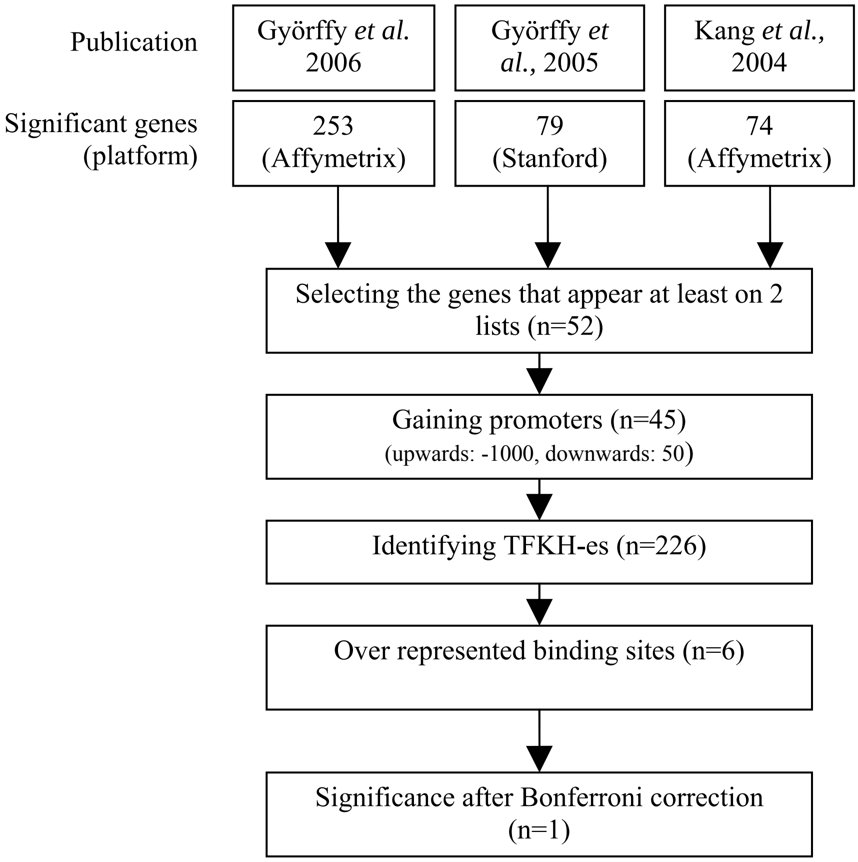
2. Defining transcription factors

Gaining sequences

First the regulation regions of the co-regulated genes had to be collected for defining the connecting places of the transcription factor. First I downloaded the proximal promoter sequences (1000 bases upwards and 50 bases downwards from the transcriptional start position) from the genomic data base using EZ-Retrieve. I saved the extracted sequencies in FASTA format and then imported to TOUCAN.

Defining the transcription factor-bonding sites

As for the next step I searched for the over-represented short DNA motifs on the promoter sequences with the help of computer programs. The TOUCAN program was used for performing the comparative promoter analysis of the selected genes. The binding sites in my sequence set were defined by the Motif Scanner equipment which constitutes part of the TOUCAN program and searches in the TRANSFAC data base. I chose the EPD human promoter group as to be the background model. In order to find the over-represented motifs I applied the statistical tool of TOUCAN for comparing data created by MotifScanner combined with the appropriate expected frequency files (human EPD). In the end Bonferroni correction was done to compensate the effect of the multiple tests (Graph 1).


Publication
Győrffy *et al.* 2006
Győrffy *et al.*, 2005
Kang *et al.*, 2004
Significant genes (platform)
253 (Affymetrix)
79 (Stanford)
74 (Affymetrix)
Selecting the genes that appear at least on 2 lists (n=52)
Gaining promoters (n=45) (upwards: -1000, downwards: 50)
Identifying TFKH-es (n=226)
Over represented binding sites (n=6)
Significance after Bonferroni correction (n=1)

Graph 1: Survey of the incoming data and the applied statistical analyses.

3. Defining the degree of antisense transcription

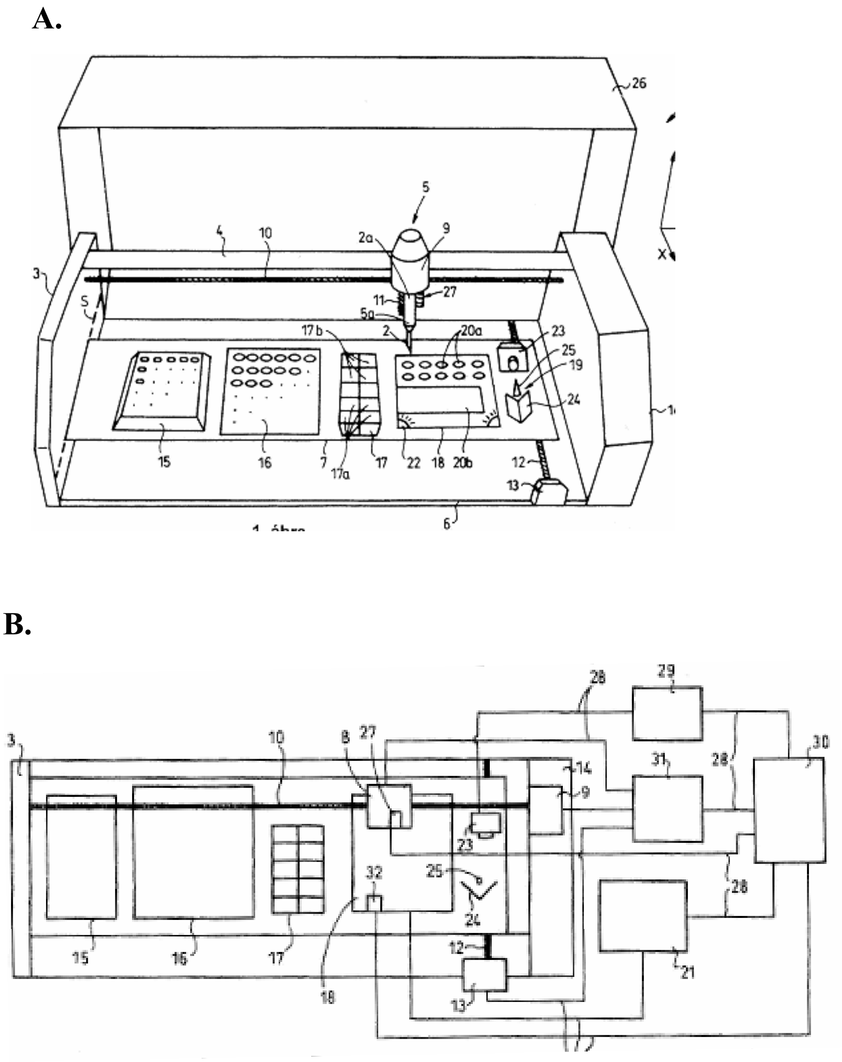
The primary objective was to identify genes where both sense and antisense expression data were available. For this the MGU74A chip version one was compared with the MGU74A chip version two with the help of the Matchprobe R library. I downloaded the complementary appropriate expression data file from the Gene Expression Omnibus (GEO) web data base. Each test sequence was subtracted and from them an oppository side complementary sequence was created. Then the identical sequence of the other chip was identified with the help of Microsoft Access. Altogether I found 8688 appropriate sequence pairs. To avoid ambiguity and to make sure that the cis antisense transcript is the exact equivalent of the sense one, I dismissed the set the sequences that pointed on different transcripts on the other MGU74A chip. That is to say I decreased my list to those sets that had had only one available match. In this manner 1182 transcripts remained with known complementary sequences and expression levels.

In my first disquisition I chose the data files published by Porter et al in 2003, since they hybridized the same samples to both versions one and two of MGU74A chips. The GEO availability numbers are: GDS614, GSD639 and GDS703. As a second set of data I selected the sequence set published by Wong et al in 2003 with the data set numbers GDS431 and GDS432. I omitted all other transcripts and focused exclusively on the above selected 1182 transcripts. Since the Affymetrix data are semi-quantitative, the relative level of the antisense partners could be taken as basis for the analysis. I set the statistical significance level to be $p<0.05$.

The Affymetrix and NetAffx Analysis Center was applied for the annotation of the transcripts. The mouse genome nomenclature was used according to the data of the Mouse Genome Informatics data base. Pearson correlation was computed for defining the ratio of the above mentioned specific sense and antisense expression. The Biobase, Annotate and Geneplotter R libraries were applied for the examination of chromosomal localization.

4. Patents and description of the automatic measuring system and positioner

The details of the invention (Patent pend: Győrffy et al, #101204) are exposed via a drawing. On the drawing, on graph 2.A the perspective picture viewed from above of the constructed form of the automated measuring system can be seen. Graph 2.B shows the rough block diagram of the measuring system and the connecting electronic control system.



Graph 2: A. Design of automatic measuring system B. Draft diagram of measuring system and its electronic control (1 automatic measuring system; 2 disposable pipette tip; 2a electric-pipette; 3 metal frame; 4 upper construction; 5 object-moving robot arms; 5a leg of pipette tips; 6 base of metal frame; 7 worksheet; 8, 9, 13, worm gear pacer; 10, 11, 12, transfer screw; 14 closed box of metal frame at right side; 15 pipette tips holder; 16 sample holder; 17 refuse bin; 18 measuring area; 19 positioning unit; 20a measurint places; 20b chip; 21 cooler/heater unit; 22 lightening unit; 23 camera; 24 mirrors; 25 reference tip; 26 cap; 27 activ pump; 28 cables; 29 sign prosessing unit; 30 control center; 31 control unit; 32 thermosensors.

With the help of the positioning unit (Patent pend: Győrffy et al, #101301) the object to be positioned, the position of the pipette tip in the present case, can be defined and set with micrometer accuracy compared to the reference tip of the fixed reference object.

The positioning unit consists of a camera, mirrors and a reference tip. The reference object is a thorn-like object, the tip of which constitutes the reference tip which can be found between two mirrors, that form an obtuse, 135° angle with each other, in the angle halfing plane of the α angle. The picture creating device, a digital camera, was fixed opposite the mirrors, being symmetric to them, in a 3-cm distance from the point of the reference tip. The mirrors positioned in an obtuse angle make the use of a second camera necessary. To ensure the appropriate intensity of light two point-like source of light can be found immediately in front of the mirrors, at the bottom of both the mirror on the right and the one on the left. The camera was connected to a sign processing unit controlled by a computer. The pipette tips are transported and moved by an object moving robot arm with the help of worm-gear pacer motors in X, Y and Z directions, based on the processed signs (photos) taken by the camera.

The positioning system is also suitable for defining the coordinates of microelectrodes used in automated pipetting equipments.

4. RESULTS

4.1 Consensus gene list connected to resistancy

Three different gene lists were combined for the comparative promoer analysis. In connection with doxorubicin resistancy, out of the total number of the found 312 specific genes 52 appeared repeatedly in at least two gene lists (Graph 1.). Out of the 52 genes, the promoter sequencies of 45 genes could be downloaded.

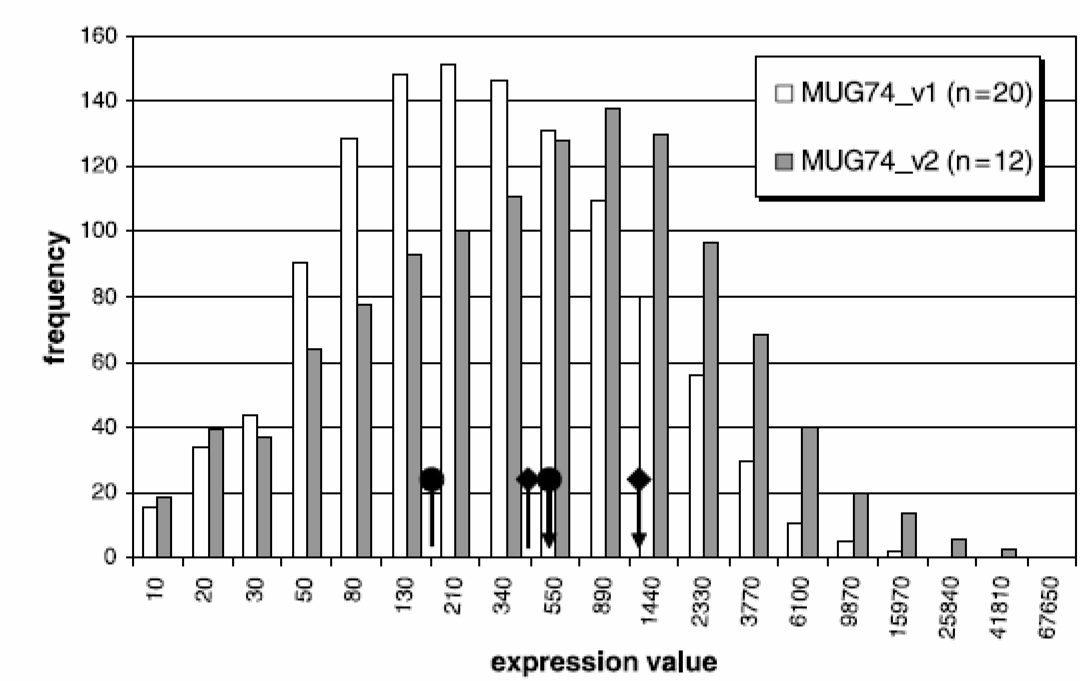
4.2 defining the common transcription factor in the obtained gene list

The comparative analysis detected E47 (p=0.00024, n=9, background frequency: 8.19E-6) as the over-expressed transcription factor of the 45 genes with available promoter sequency.

This transcription factor-binding site appeared on 9 different genes, among them on SKP2 gene with 2 binding sites. The E47 TFKH consensus sequence is: RCAGNTG (TRANSFAC availability number: R02139).

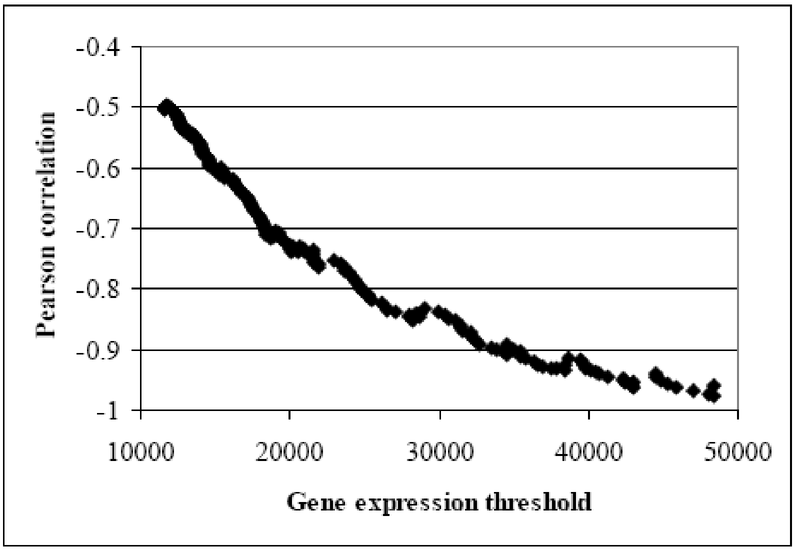
4.3 Defining the degree of antisense transcription

The expression differences of the MGU74A version one and version two chips were compared in Porter et al`s database. In the investigated transcripts a 43% general antisense expression was found compared to the sense expression (Graph 4). The histogram shows the expression levels of both chips showing that the antisense expression levels are smaller than the sense ones (Graph 4). The average antisense expression in relative normalized unit on the MGU74A version one chip was 525.1, whereas it was 1219.1 (t-test: $p<0.0001$ Graph 3) on the sense chip.



Graph 3: Histogram of expression values of all 1182 genes due to 20 antisense (MGU74A_v1) and 12 sense chips (MGU74A_v2). You can see average expreesion values of MGU74A chips (arrows: average of MGU74A_v1 is 525,1; average of MGU74A_v1 is 1219,1; p>0,001. Average without arrows v1: 196,4; v2: 414,6).

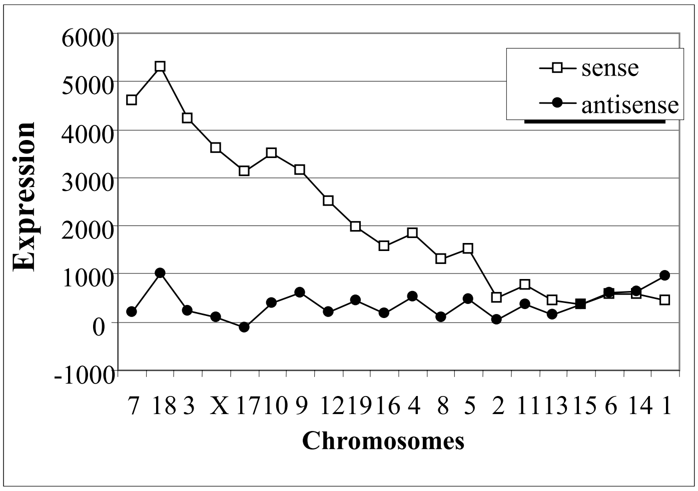
During examination of dataset published by Wong et al I have found gene expression in both directions. Average antisense expression was found to be 516 (in normalized relative unit) and average sense expression level was 1688. At extremely high expression values the frequency of sense genes is bigger than antisense genes (Graph 4).



Graph 4: This graph shows correlation between sense and antisense gene expression above a relative gene expression threshold. I have set threshold value due to higher expression for both sense and antisense transcripts. Pearson-correlation is At higher expression values Pearson-correlation approaches -1 confirming strong inverse correlation between sense and antisense gene expression.

I selected those transcripts which had significant higher sense-antisense ratio. So I found transcripts with changed antisense expression which had been related to the MHC study.

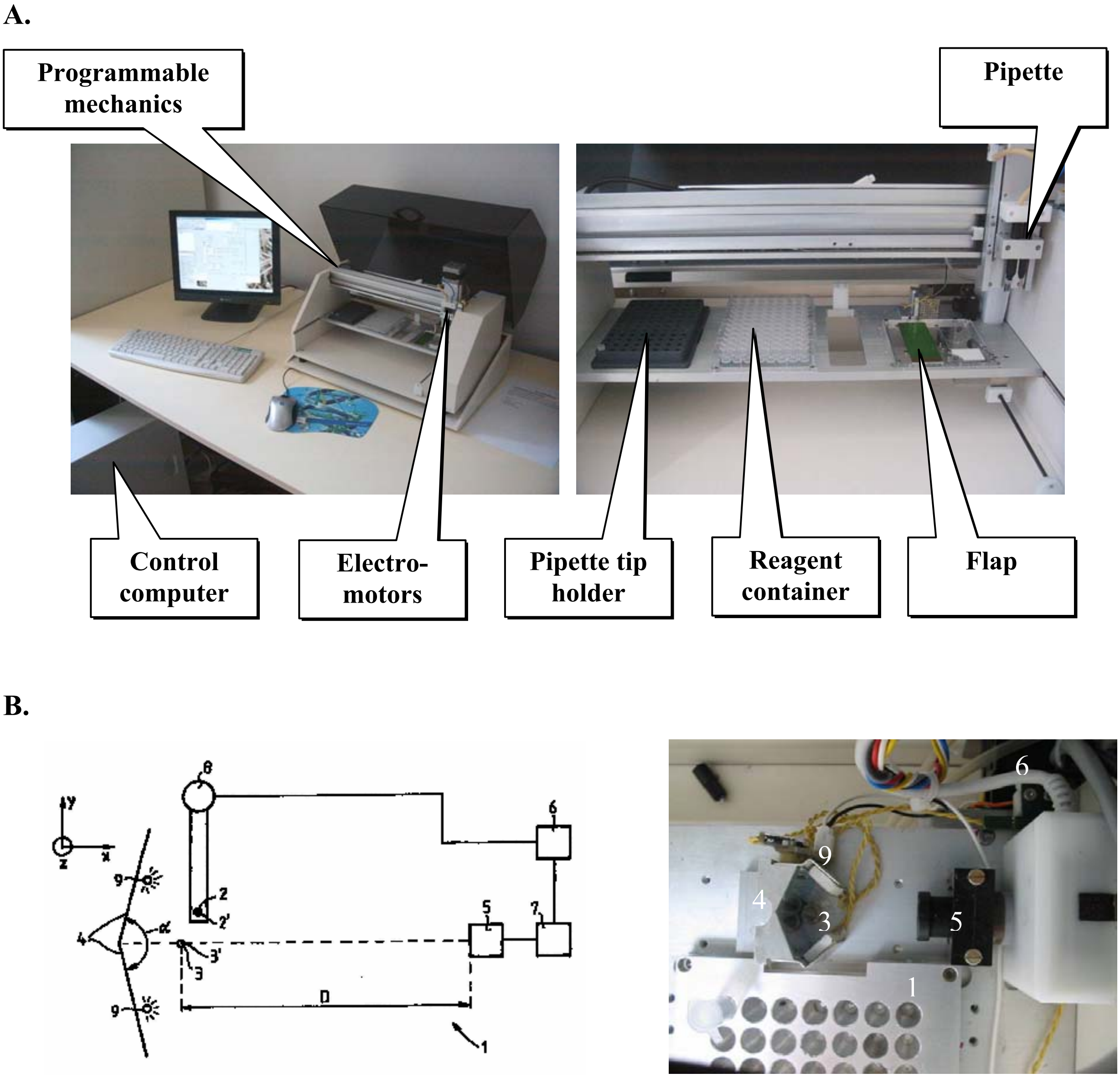
I drew chromosomal locations of every single transcripts to examine them and also the relative transcription. I found higher relative antisense expression in 1 and 14. chromosomes. Sense and antisense transcription was approximately the same in the 6. and 15. chromosomes. In all other cases the sense transcription was higher than the antisense (Graph 5).



Graph 5: This graph shows the average sense and antisense expression of every chromosome. Chromosomes are listed by differences between their sense and antisense expression.

4.1 Automatic measuring system and positioning method

My invention is an automatic measuring system applicable for electrochemical and biochemical measurements which is suited for containing samples with one or more sample holders, with measuring surface, sample receptacle, one or more measuring instruments and it also has an object holding and moving instruments for sample reception (Graph 6-A,B).



Graph 6: A. Design of automatic measuring system. B. Draft and real image of positioning unit from above.1. Distance between the camera and the reference tip (D); 2. and 2` the pipette tip and its top; 3. and 3` reference object and its tip; 4. mirrors at α angle related

to each other; 5. ccd camera; 6. control unit; 7. sign processing unit (PC); 8. arm of swelter; 9. lightning units on the bottom of mirrors.

5. CONCLUSIONS

Definition of oligo regulating factors

- I determined a list of consensus genes which can be related to doxorubicin resistance. In the course of analyses of previous relevant studies I identified 52 genes which can be responsible for the development of resistance due to their biological function.
- The overrepresentation of E47 transcription factor has been found on the promoter sequences of genes in association with doxorubicin resistance. The E47 transcription factor can be the key factor of the regulating system behind the development of doxorubicin resistance.

The examination of oligo sequence direction and its regulating functions

- My studies and the outlined data have shown that the antisense coexpression and inverse expression are frequent and inner features of the mouse genome. It can be extrapolated by the extension of the examination that the general antisense expression regarding all mouse genes can be 43%.
- I have found inverse correlation between sense and antisense mRNS expression at high sense (and antisense) expression values.
- When examining the sense and antisense rates I have found differences among certain chromosomes. I have detected the lowest antisense expression level on chromosome 9 and 17, the highest ones on chromosomes 1 and 14. The level of antisense expression was low on X chromosome.

Development of oligo spotting technology

- I have invented an automatic measuring system suitable for preparing and accomplishing molecular biological reactions and measuring while it can store the basic materials, reagents and tools necessary for the examinations at a desirable temperature and in a closed field. A surface layer can be grown on the firm carrier with the tool.

- I have developed a positioning method and built a positioning unit able to 0,1 mm accurate positioning of the disposable commercial 0,1-20 μl pipette tips.

6. SUMMARY

6.1 The investigation of the gene expression trans regulating elements

Behind the evolution of genes possessing an identical expression profile a regulating mechanism has to be found which is responsible for the simultaneous co-expression of these genes. My basic hypothesis is that in a given experiment the same transcription factors can be responsible for the regulation of genes possessing the identical expression profile. The process was investigated through a clinical problem, namely when resistance developed against the applied doxorubicin treatment during cancer chemotherapy. During the research project it was shown that correlation can be assumed between the detected E47 transcription factor and medicine resistanc. These results bring new light to E47 that does not play a role only in cell proliferation as it had been assumed earlier.

To this effect E47 can become a new target to counter tumour doxorubicin resistancy.

6.2 The investigation of cis gene expression regulation

It is more and more obvious that in the human and other eucariotic genomes the antisense transcription is wide spread. Based on the results of the last years it is unambiguous that antisense transcription bears a key role in the evolvement of certain human diseases. Although the sense-antisense pairs have been widely studied more thoroughly, we know little about the antisense expression of the known genes.

During the research I examined and compared the sense and antisense expressions in two papers, investigating 1182 mice transcripts, in order to determine the frequency of antisense transcription and to find signatures characteristic of antisense transcription. In both of my investigations I found antisense transcription compared to sense transcription to be of the same frequency but of a smaller degree. According to my investigations the antisense co-expression and inverse expression is a general feature of the mouse genome. Based on my results the general antisense expression can be estimated to 43% considering all mouse genes.

In my second analysis I found a direct inverse correlation between the sense and antisense figures at large expression figures. In this case, at very high relative figures (around 40 000 relative expression unit) the Pearson correlation coefficient approximated -1. Transcripts possessing high inverse expression rate correlated with the originally examined MHC II. transactivator genes. Consequently my results support the regulating functions originating from the inverse nature of the antisense transcription.

In both investigations the sense-antisense transcription rate showed an altering degree on the various chromosomes. The antisense expression is higher than the sense one on the 1st and 14th chromosomes, while very low antisense transcription was found on the X chromosome, this fact coinciding with the earlier examinations. My data investigating the chromosome specific signatures underline the potential role of the sense-antisense rate in the process of expression of alleles.

6.3 The development of oligo spotting technology

Until now there has not been available any device where different drops of liquid could have been placed on an optional flat surface. It is an element of key importance during the making of DNA chips to position the oligonucleotids fixed on microarray according to an appropriate array and program and subsequently to do the coordinated implementation of the proper hybridizing reactions. Due to the high degree of sensitivity of the examinations it is indispensable to prevent the compounds from injuries. Besides this it is also crucial to ensure closed, sterile circumstances for making chips if this may be the case.

During the research I created an automated pipetting measuring system with the help of which pipetting on an arbitrary surface and performing electrochemical and molecular measurements there can be done cheaply and simply. The automated measuring system is apt for preparing and performing molecular biological reactions and measurements while it stores the resources, reagents and the materials constituting the objects of the given investigation at the desired temperature and in closed space. The device is capable of growing the surface layer, testing small amounts of poisonous reaction components, fluorescent measuring or performing measurements connected to electrochemical or biochemical reactions that require sensitive to light components and it is also apt for making examining tools and DNA-microarrays.

The automated pipetting device can be arbitrarily programmed, the liquid reaction components can be positioned on its work plate to an optional place and in optional order within a 0.1-20 µl range with 0.1 µm accuracy.

This system uses disposable pipette tips that are available cheaply and easily in trade flow. The above mentioned accuracy can be attained only if the position of all the pipettes can be determined in the three-dimensional space with 0.1 μm accuracy. I chose that method of positioning the pipette tips where I transport the pipette tips to a known reference summit which I set to be the origo coordinate point and I correlate to it with a computer controlled picture processing system. Thereafter the disposable pipette tip can be motioned anywhere by a computer controlled robot arm.

Due to the complexity of the automated measuring system it is apt for several more types of meaurements other than the ones listed here. Further measuring devices or systems can be built into it optionally, via which different specific measuring tasks can also be designed.

7. PUBLICATIONS

Papers: 9	Impact factor: 11,029	Submitted patents: 2

Publications which underlie and are related to this Ph.D. work:

1. **Győrffy A**, Vasarhelyi B, Szoke D és mtsai: Comparative promoter analysis of doxorubicin resistance-associated genes suggests E47 as a key regulatory element. **Anticancer Res** 2006;26:2971-2976. *(IF: 1,479)*
2. **Győrffy A**, Surowiak P, Tulassay Z, Gyorffy B: Highly expressed genes are associated with inverse antisense transcription in mouse. **Journal of Genetics** *(IF: 0,528, in press)*
3. **Győrffy A**, Z Tulassay, B Gyorffy: Computational Analysis Reveals 43% Antisense Transcription in 1182 Transcripts in Mouse Muscle. **DNA Sequence** 2006:17:422-430. *(IF: 0,569)*
4. **Győrffy A,** Z Baranyai, A Cseh, Gy Munkacsy, F Jakab, Z Tulassay, B Gyorffy: Promoter analysis suggest the implication of NFkB/C-REL transcription factors in biliary atresia. **Hepato-Gastroenterology** *(IF: 0,756, in press)*
5. Szőke D, **Győrffy A**, Surowiak P, Tulassay Z, Dietel M, Györffy B: Identification of consensus genes and key regulatory elements in 5-fluorouracil resistance in gastric and colon cancer, **Onkologie**, *(IF: 1,724, in press)*

Publications in Hungarian:

6. **Győrffy A**, Makai D, Gyorffy B, Harsanyi G, Tulassay Z: Microelectrodes and their application in diagnostic medicine. **Orv Hetil** 2006;35:1703-1708.
7. **Győrffy A**., Gyorffy B., Molnar B., Tulassay Z: Hybridization and their application in the DNA array technology. **Orv Hetil** 2005;27:1447-1452.
8. Győrffy B, **Győrffy A**, Tulassay Z: The problem of multiple testing and solutions for genome-wide studies. **Orv Hetil** 2005;12:559-563.

Additional publications not related to the presented theses:

9. Kocsis I, Vasarhelyi B, **Győrffy A**, Gyorffy B: Reanalysis of genotype distributions published in "Neurology" between 1999 and 2002. **Neurology** 2004;63:357-358. *(IF: 5,973)*

Patents:

1. Positioning method and unit, especially for positioning pipette tips Győrffy A., Győrffy B., Virág T., Molnár B., Szabó A., Tulassay Z., Tulassay T., (P0500670-2005, in progress)
2. Automatic measuring system especially for electrochemical and biochemical measurements Győrffy A., Győrffy B., Virág T., Molnár B., Szabó A., Tulassay Z., Tulassay T., (P0500671-2005, in progress).

Conference performances:

1. 2007. Conference of the Hungarian Society of Gastroenterology , Diagnostic gene expression signatures to predict the chemotherapy resistance against oral and intravenous 5-fluoruracil. Munkácsy Gy., Győrffy B., Baranyai Zs., Győrffy A., Jakab F., Tulassay Zs
2. 2007. PhD Conference, Development of Diagnostic gene expression signatures to predict the chemotherapy resistance against 5-fluoruracil. Munkácsy Gy, Győrffy A, Baranyai Z, Jakab F, Tulassay Z, Győrffy B,
3. 2006, isse2006, st. Marienthal, electrochemical dna Sensor Development for Diagnostic Application D Makai,. A. Győrffy, G Harsányi, Z. Tulassay
4. 2006, PhD Conference, Budapest, (II. award): Mikroelectrodes in medical diagnostics. A. Győrffy, D. Makai, B. Győrffy, G. Harsányi, Zs. Tulassay
5. 2006 Kandó Kálmán Academy, Budapest, Hibridization and its thermodinamical basics, invited lecturer.
6. 2005, 25-276th International congress of the worldwide hungarian medical academy (WHMA) Budapest, Genom-wide antisense transcription is suggested by the analysis of 1182 mouse transcrpits A. Győrffy , B Győrffy
7. 2004, UEGW, Prague, Development of a new electrochemical chip technology for gastrointestinal biopsy specimen cDNA analysis Győrffy A., Virág T., Győrffy B., Szabó A., Tulassay T., Patócs A., Molnár B., Tulassay Z.
8. 2004, Conference of the Hungarian Society of Gastroenterology, Balatonaliga, DNA immobilisation and detection, Győrffy A., Molnár B., Tulassay Z.
9. 2004, Hungarian innovation exhibition, New DNA-preparing technology and demonstration of instrument., A. Győrffy
10. 2004, Semmelweis University PhD Conference, Budapest, Summarising Chip hibridisation technology. A.Győrffy
11. 2003, ISSE 2003, Mátrafüred, DNA chip with electronically accelerated processes A. Győrffy, H. Sántha
12. 2002, Polytronic conference, New methods of DNA oligomer - binding and detection in the DNA microarray technology A. Győrffy, A. Patócs, H. Sántha, Rácz K., G. Harsányi