



ELSEVIER

Available online at www.sciencedirect.com

Computer Physics Communications ●●● (●●●●) ●●●-●●●

Computer Physics
Communicationswww.elsevier.com/locate/cpc

Analysis of DNA-chip and antigen-chip data: studies of cancer, stem cells and autoimmune diseases

Eytan Domany

Department of Physics of Complex Systems, Weizmann Inst. of Science, Rehovot 76100, Israel

Abstract

Biology has undergone a revolution during the past decade. Deciphering the human genome has opened new horizons, among which the advent of DNA microarrays has been perhaps the most significant. These miniature measuring devices report the levels at which tens of thousands of genes are expressed in a collection of cells of interest (such as tissue from a tumor). I describe here briefly this technology and present an example of how analysis of data obtained from such high throughput experiments provides insights of possible clinical and therapeutic relevance for Acute Lymphoblastic Leukemia. Next, I describe how gene expression data is used to deduce a new design principle, “*Just In Case*”, used by stem cells. Finally I briefly review a different novel technology, of antigen chips, which provide a fingerprint of a subject’s immune system and may become a predictive clinical tool. The work reviewed here was done in collaboration with numerous colleagues and students.

© 2005 Elsevier B.V. All rights reserved.

PACS: 87.10.+e; 87.90.+y

Keywords: Gene expression; DNA microarray; Leukemia; Stem cells; Antigen chip

1. Introduction

I present here a brief and oversimplified description of very complex processes. For a more detailed review see [1]; readers interested in learning the subject are referred to two excellent textbooks [2,3].

DNA is a one-dimensional molecule, made of two complementary strands, coiled around each other as a double helix and held together by hydrogen bonds that connect a linear sequence of complementary pairs

of bases. There are four kinds of bases, denoted by C, G, A, T; the bound pairs are G–C (three hydrogen bonds) and A–T (two). DNA resides in the nucleus of (eucaryotic) cells.

A *gene* is a segment of DNA, which contains the formula for the chemical composition of one particular protein. The *genome* contains the collection of all the genes that code for all the proteins that an organism needs and produces. A *gene is expressed* in a cell when the protein it codes for is actually synthesized. The human genome has between 20 000 and 30 000 genes. Even though each cell contains the entire genome, different genes are expressed in different cell types.

E-mail address: eytan.domany@weizmann.ac.il (E. Domany).

Transcription and translation. Synthesis of proteins takes place at the *ribosomes*, huge complexes that reside in the cytoplasm; the information encoded on the DNA is transferred from the nucleus to the ribosomes by a molecule called messenger RNA (mRNA). When a gene is expressed, precise copies of the information written on one of the DNA strands are made, in the form of linear mRNA molecules, that leave the nucleus and diffuse through the cytoplasm. The process of copying DNA onto mRNA is called *transcription*. Subsequently a ribosome reads the message from the mRNA and *translates* it into a sequence of amino acids, added one at a time, comprising the corresponding protein. When many copies of a certain protein are needed, the cell produces many copies of corresponding mRNA molecules, which are “read” by several ribosomes.

A particular cell may need a large number of some proteins and a small number of others. That is, every gene may be expressed at a different level. This leads us to the *basic paradigm of gene expression analysis*: The “*biological state*” of a cell is reflected by its *expression profile*: the expression levels of the entire genome. These, in turn, are reflected by the concentrations of the corresponding mRNA molecules. By “*biological state*” I mean whether the cell is normal or malignant? is it starving? To which tissue type it belongs? Is it undergoing cell division? etc.

A *DNA chip (microarray)* is a device that measures the concentrations of thousands of different mRNA molecules [4,5] The state-of-the-art Affymetrix [6] microarray, U133SP2, measures the expression levels of 54 675 probe sets (that represent genes and other transcribed bits of DNA). The underlying physical process is *hybridization*: billions of bits of DNA are printed on the chip and serve as *probes*, while mRNA molecules extracted from the cells that are studied play the role of *targets*; they are brought into contact with the chip and if they find a matching probe, they bind to it. Fluorescent markers are attached to the bound mRNA, illuminated by a laser, and the light intensity from a pixel measures the number of targets that have stuck, which is proportional to the concentration of these mRNA in the studied tissue. A typical experiment provides the expression profiles of $N_s \approx 10\text{--}100$ tissue samples, over $N_g \approx 10\,000$ genes, summarized in an $N_g \times N_s$ *expression table*: E_{gs} is the expression level of gene g in sample s .

2. Acute Lymphoblastic Leukemia (ALL)

Leukemia is a malignancy of the hematopoietic system; Hematopoietic Stem Cells (HSC) differentiate gradually as they divide and proliferate, becoming either mature Lymphoid cells (e.g., B and T cells) or Myeloid (e.g., platelets, red blood) cells. During differentiation genetic alterations—such as point mutations or translocations—may occur, which destroy the (normally very tight) control over cell division. If the genetic disaster occurred in a precursor of the Lymphoid cells, the disease is called ALL; the resulting malignant cells (blasts) (a) proliferate at a very high rate, and (b) differentiation stops, the blasts never mature into functional normal cells. These blasts overcrowd the organism, depriving normal cells of their life-space and the patient loses his immune system.

Our work [7] focused on the effects of a particular translocation, denoted $t(4, 11)$, of a gene called MLL, located on chromosome 11, with a partner gene from chromosome 4. When a *translocation* occurs, DNA from one chromosome is joined to DNA from another, possibly creating a chimeric protein, which has the tail of one “normal” protein fused to the head of another. MLL translocations are implicated in 80% of infant ALL cases and in the majority of therapy-related secondary leukemia. The chimeric MLL protein is an *oncogene*: it induces cancer by activating some target genes and suppressing others. The protein content of the cell changes in a way that causes cancer. Our *aim* was to identify these target genes of the $t(4, 11)$ MLL translocation. 14 ALL patients with $t(4, 11)$ were identified, and 13 without. The expression levels of 12 000 genes in blast samples, taken from these 27 patients, were measured; 3064 of these passed certain quality-control filters. The Wilcoxon rank-sum test identified genes differentially expressed between the two groups of samples. The problem of multiple comparisons was controlled by the False Discovery Rate (FDR) method [8]; at an FDR of 10% we found 237 such genes (e.g., the expected number of false positives was ≈ 24). It became clear, however, that most probably, not all 237 genes are genuine MLL targets! the reason is that in addition to the presence/absence of the MLL translocation, the blasts taken from the two groups of ALL patients differ by a second factor: early versus late differentiation stage. In the MLLs arrest of differentiation is believed to oc-

cur at an early stage, whereas in the ALLs without the MLL translocation, late differentiation is more likely. Hence for some of the genes, the difference of expression levels between the two groups could reflect only the difference in the stage at which differentiation came to a halt, and *not* the transcriptional activity of the MLL oncogene. By identifying a subgroup of 4 ALLs without the translocation that are known to be early differentiators, we succeeded to sort out [7] which are the differentiation-sensitive genes and narrowed the list of candidate MLL targets down to 43 genes, 18 of which were overexpressed in the MLLs. The differential expression of these 43 genes was confirmed by checking their expression data of two other groups [9,10]. Future experiments will establish which of these 43 genes play a central role in generating ALL; these will be candidate targets for drug therapy. Surprisingly, we also discovered that contrary of “common wisdom”, (a minority of) the translocated MLL blasts can be late differentiators.

3. Just In Case: a design principle used by stem cells

Most cells of a multicellular organism are mature, fully differentiated, and have a fixed function (a B cell cannot become a neuron). Their division rate is slow, the number of divisions is limited. The lifetime is finite; most mature cells die; but if these are essential for life, there must be a constant resupply. *Stem cells* are the source of this incessant “flow” of cells; they are immortal, can undergo an unlimited number of divisions and their number is kept constant by a mechanism of self-renewal. When stem cells divide, half of the daughter cells remain stem cells and the other half progress and differentiate into various kinds of mature tissue.

Fully differentiated cells of distinct lineage have different functions and hence need different proteins. A dividing stem cell does not “know” the fate designated to each of its progenies. *What is the preferred strategy of gene expression that such a cell will follow?* One can think of two extremes; the parsimonious *Just In Time* strategy [11] and its opposite, the wasteful *Just In Case* [12]. Under the former, stem cells do not express genes that are used by mature tissue; as differentiation progresses towards a particular fate, genes

will start to be expressed just at the right time, when they become needed. The opposite strategy calls for expression of a large number of genes, to have them around and ready, just in case a cue arrives for commitment to a cell type that in fact needs them. Genes are “shut down” as differentiation progresses and it becomes clear that they are not needed for the actual target tissue. This strategy is more wasteful, but it gives the differentiating cell a faster response time to such cues.

To resolve the strategy, we measured gene expression profiles of three types of human cells: embryonic stem cells (with the widest range of potential mature targets); hematopoietic stem cells (whose range of end products is limited to a smaller number of such targets) and of fully differentiated mature cells (blood). On the basis of cluster analysis [13] of the expression profiles of thousands of genes over these three types of cells we concluded [12] that embryonic stem cells *mix the two strategies*, with a significantly larger group of genes behaving according to *Just In Case*. We repeated the analysis using embryonic stem cells with keratinocytic stem cells and mature keratinocytes (skin) and reached a similar conclusion; hence the *Just In Case* strategy seems to be universal, used along distinct differentiation paths.

4. Antigen chips

The immune system recognizes, marks and destroys invaders. Invading organisms (e.g., bacteria) have membrane proteins that are recognized by the system as non-self. Molecules that trigger an immune response and are eventually recognized as foreign are called *antigens*. They are recognized by proteins called *antibodies*, which bind to the antigen and trigger a chain of complex events that cause the eventual destruction of the invader. The immune system “learns” the shape and chemical composition of an antigen to which it has never before been exposed, by generating new antibodies that bind to the antigen. The system “remembers”; next time the same antigen appears, it is greeted by a well prepared set of antibodies. Hence the antibodies that each of us carries around in his serum reflects our “biological history” [14]. Since the number of possible antibodies is enormous (probably many millions), it is impossible to measure the

concentration of all possible antibodies of an animal. Can one obtain a fingerprint of the immune system by measuring a much smaller set of numbers?

To answer this question, we developed *Antigen chips* [15]: one prints on a glass plate about 300 spots, each containing one particular antigen. We selected antigens that are either known to play a role in the autoimmune disease we set out to study (diabetes) or were of interest on some other ground. Serum taken from a subject is brought into contact with the spots; it contains an enormous number of different antibodies, at different concentrations. Binding is not 100% specific; antibodies bind, with varying reactivities, to every antigen. Some spots probably have several types of antibody bound to them and some are relatively free of bound antibodies. After hybridization and washing only antibodies with high concentration and reactivity with the antigen on a spot stay bound. Next, the chip is exposed to molecules that recognize the bound antibodies—these are then fluorescently labeled, the chip is scanned by a laser, and the intensity of fluorescent light emanating from each spot measures the number of antibodies bound to the antigen on that spot. The resulting 300 numbers constitute the *reactivity profile* of the subject whose serum was tested. Previously [16] we tested 90 antigens; not a single one could separate diabetics from healthy subjects in a statistically significant way, but the reactivity profiles of *groups* of antigens (identified by advanced clustering techniques [13,17]) did assign diabetics and healthy subjects to different clusters.

Predicting outcome of treatment. In recent work [18] we studied 20 genetically identical Non-Obese Diabetic (NOD) mice. Mice of this strain spontaneously contract diabetes at the age of 11 weeks. If injected with Cyclophosphamide (CY), half the mice develop CY-induced accelerated diabetes (CAD) at the age of 6 weeks, while the other 10 do not become diabetic at all; they are immunized by the treatment. The question we posed was—can one, on the basis of the reactivity pattern of blood taken *prior to injection of CY*, distinguish between the two groups—CAD versus immunized? No antigen separated the mice on its own. We picked the 10% (27 antigens) that best separate the mice and clustered all 20 mice. on the basis of these 27 antigens. Two clear, statistically significant clusters were found; one rich in CAD and the other in immunized mice. This experiment demonstrates that

individual differences between “genetically identical” mice suffice to give rise to different responses of the immune system to treatment; these differences are inscribed *prior* to the introduction of the treatment, and can be picked up from the reactivity patterns of the antigens placed on our chip.

5. Summary

Experimental methods used in biology move towards miniaturization of the measuring devices, which allows massive parallelization of data acquisition and generates large volumes of measured data. The main advantage of these high-throughput techniques is also their weakness; the need to sort out the relatively few aspects of the measurements that are relevant to the biological questions of interest pose a computational challenge. The examples reviewed above demonstrate how this type of analysis yields clinically relevant information (on leukemia), basic insights into design principles (of stem cells) and potentially opens a field of predictive immune fingerprinting.

Acknowledgements

I thank my students G. Getz, O. Ravid-Amir, M. Golan-Mashiach, O. Barad, G. Hed, G. Elizur and P. Hagedorn, whose contributions I described here, and my colleagues and collaborators; E. Canaani, T. Rozovskaia, J.-E. Dazard, G. Rechavi, D. Givol, F. Quintana and I. Cohen, for so generously sharing their data and knowledge. The work described here was supported by the Israel Academy of Sciences, the Minerva and Ridgefield Foundations, the Germany–Israel Science Foundation (GIF) and the European Community’s Human Potential Programme under contract HPRN-CT-2002-00319, STIPCO.

References

- [1] E. Domany, *J. Stat. Phys.* 110 (2003) 1117.
- [2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *Molecular Biology of the Cell*, Garland Publishing, New York, 2002.
- [3] J.L. Gould, W.T. Keeton, *Biological Science*, W.W. Norton and Co., New York, 1996.
- [4] A. Schulze, J. Downward, *Nature Cell. Biol.* 3 (2001) 190.
- [5] See, Special supplement, *Nature Genet.* 21 (1999) 1.

- [6] See <http://www.affymetrix.com>, for information.
- [7] T. Rozovskaia, et al., Proc. Nat. Acad. Sci. 100 (2003) 7853.
- [8] Y. Benjamini, Y. Hochberg, J. Royal Stat. Soc. Ser. B 57 (1995) 289.
- [9] E. Yeoh, et al., Cancer Cell 1 (2002) 133.
- [10] S.A. Armstrong, et al., Nat. Genet. 30 (2002) 41.
- [11] S. Kalir, et al., Science 292 (2001) 2080.
- [12] M. Golan-Mashiach, et al., FASEB J., in print.
- [13] M. Blatt, S. Wiseman, E. Domany, Phys. Rev. Lett. 76 (1996) 3251.
- [14] I.R. Cohen, Tending Adam's Garden: Evolving the Cognitive Immune System, Academic Press, New York, 2004.
- [15] G. Elizur, M.Sc. Thesis, Weizmann Inst. of Science, 2004.
- [16] F.J. Quintana, G. Getz, G. Hed, E. Domany, I.R. Cohen, J. Autoimm. 21 (2003) 65.
- [17] G. Getz, E. Levine, E. Domany, Proc. Nat. Acad. Sci. 97 (2000) 12079.
- [18] F.J. Quintana, P. Hagedorn, G. Elizur, Y. Marbel, E. Domany, I.R. Cohen, Proc. Nat. Acad. Sci. 101 (2004) 14615.