STOCHASTIC MODELS IN BIOLOGICAL SCIENCES
BANACH CENTER PUBLICATIONS, VOLUME 80
INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES
WARSZAWA 2008

THERMODYNAMICS OF DNA MICROARRAYS

ENRICO CARLON
Interdisciplinary Research Institute of Lille
École Polytechnique Universitaire de Lille (Polytech'Lille)
Cité Scientifique, F-59652 Villeneuve d'Ascq, France
E-mail: enrico.carlon@polytech-lille.fr

Abstract. DNA microarrays have been widely used in molecular biology laboratories. The main current application of these devices is the determination of the gene expression level for thousands of genes simultaneously. Here we review a recently introduced physical model for hybridization (i.e. the binding of complementary DNA strands) in Affymetrix arrays and compare it to experimental results. The experimental data follow rather well the microscopic model and the approach offers several advantages compared to traditional methods of microarray data analysis.

1. Introduction. A gene is a small fragment of DNA which contains the code for the synthesis of a specific protein (for simplicity we limit ourselves here to discuss the protein coding genes, as there are indeed genes that do not encode for a protein). The process leading to the synthesis of a protein consists of two steps: the gene is first transcribed into messenger RNA (mRNA) which is then translated into protein.

The 30,000 human genes are not expressed, i.e. transcribed into mRNA, at all times. Some genes are tissue specific in the sense that they are expressed only in some given tissues, while others are more ubiquitous and are expressed in all tissues. For instance: pancreatic cells specialize in the synthesis of proteins as trypsin, which is an enzyme involved in the digestion and a main component of the so-called pancreatic juice. The associated gene is highly expressed in pancreatic cells, while is typically not expressed in cells of other tissues. This information alone suggests that trypsin is a protein which is involved in some basic pancreatic function. Knowing when, in which amount and in which tissue a given gene is expressed suggests the functions of the corresponding protein. This is a very valuable information to understand the function of our genes: currently the function of only about 20% of the 30,000 human genes has been identified.

2000 *Mathematics Subject Classification*: Primary 92E20; Secondary 80A10.
Key words and phrases: DNA microarrays, DNA hybridization.

The paper is in final form and no version of it will be published elsewhere.

DNA microarrays are devices which, in principle, can measure the gene expression level of thousands of genes simultaneously. They are made of a small glass plate containing single stranded DNA fragments anchored at the surface. A microarray experiment consists of the following steps: first a class of cells is extracted from a tissue and the total mRNA from those cells is isolated and marked with fluorescent molecules. The solution obtained is then put in contact with the microarray. If a given sequence floating in solution (known as a *target*) meets its complement sequence anchored at the glass surface (the *probe*) it binds. The binding of complementary probe-target sequences is called hybridization. As the target strand is labeled with fluorescent molecules the fraction of hybridized probes can be measured by optical methods. Fig. 1(left) shows an image of an Affymetrix (a leading company in the DNA microarrays market) array.

Here we will review some recent results of an analysis of DNA microarray data using physical-chemistry models. We will focus on microarrays produced by Affymetrix. These arrays have two special features compared to all other arrays: 1) Multiple probes per gene are used 2) for each probe perfectly complementary to DNA in solution there is an additional probe with a sequence differing by a single nucleotide. Thus, one distinguishes between perfect matching (PM) probes and mismatching (MM) ones.

In the next sections we describe the model and the analysis of the experimental data. Several other models have been proposed in the recent literature (for a review see [Halperin] and references therein). The main features of the model presented here are 1) the use of the hybridization free energies for RNA/DNA duplexes, 2) the inclusion of the mismatches in the analysis, 3) the use of very few free fitting parameters and 4) the inclusion of the hybridization between partially complementary strands in solution.

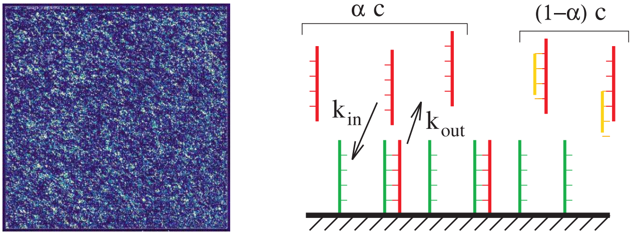

Fig. 1. Left: Fluorescent image of a scanned Affymetrix arrays. The image is of an array containing a grid of $640 \times 640 = 409,600$ different probes on a surface of 1.28×1.28 cm^2. In the latest Affymetrix chips (2006) the same surface area can contain up to 6.4 millions different sequences. Right: Hybridization model developed in [Carlon]. 1) probes with a specific DNA sequence hybridize with complementary molecules in solution with attachment/detachment rates k_{in} and k_{out}. 2) Target molecules can partially hybridize in solution leading to an effective reduction of the target concentration to a factor αc.

2. The extended Langmuir model. As the Langmuir model in its basic form does not seem to fit sufficiently well the experimental microarrays data (see [Carlon]) we have modified it to include the possibility of hybridization between targets in solution.

According to this *extended* model the fluorescent intensity measured for a given probe is [Carlon]

$$I = I_0 + \frac{A\alpha c e^{\beta \Delta G}}{1 + \alpha c e^{\beta \Delta G}} \tag{1}$$

where c is the target concentration, ΔG is the hybridization free energy, $\beta = 1/RT$ the inverse temperature (R the gas constant), A a scale factor and I_0 the non-specific background signal. The concentration c is a measure of the gene expression level that one wants to determine. The difference with the basic Langmuir model is the inclusion of a factor α, which takes into account of the effective reduction of the target molecules in solution [Carlon], as illustrated in Fig. 1. In the model we handle PM and MM signals on the same footing: perfect matches hybridize more efficiently to the probes than mismatches and this difference results in a different hybridization free energy ΔG. The latter is calculated from experimental data for RNA/DNA hybrids taken from [Sugimoto] using the nearest neighbor model. The hybridization in solution is taken into account in a kind of mean-field approximation and the sequence-dependent factor α contains two adjustable parameters [Carlon]. Two other fitting parameters in Eq. (1) are A and β: the parameter A sets the overall level of fluorescence, given in arbitrary units by Affymetrix and β is an inverse effective temperature. The latter is higher than the experimental temperature suggesting that the interaction strength in hybridization with surface-bound DNA, as in Affymetrix microarrays, is overall lower than the corresponding interaction between two strands hybridizing in solution. For a detailed discussion of this effect see [Carlon].

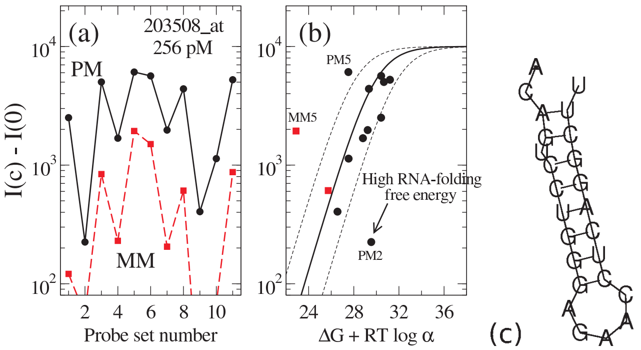

Fig. 2. (a) Plot of the background-subtracted intensity of the Affymetrix spike-in experiment for perfect matches (PM) and mismatches (MM) probes of the probe set 203508_at (corresponding to the gene 1B of the tumor necrosis factor superfamily), which in the experiment is spiked at 256 pM. (b) The same data plotted as function of the hybridization free energy according to Eq. (1) (similar plots can be produced with a free software SpikeLI, available under the Bioconductor project [SpikeLI]). The solid line is the predicted Langmuir isotherm with c = 256 pM. The dotted lines correspond to concentrations four times higher and smaller. (c) Minimal free energy configuration for the target sequence in solution which is complementary to probe 2. The low signal PM2 observed in (b) is therefore very likely caused by the folding of the sequence in solution.

3. Data analysis. We have analyzed freely available Affymetrix data of two Latin square experiments. In these experiments some targets are spiked-in at a known concentration ranging from 0.125 up to 1024 pM (picomolar), covering all the concentrations of biological interest. As c is known in Eq.(1) we have used the data to fit the other four global fitting parameters. Fig. 2 gives an example of the signal measured from a probe set, ie the collection of all probes PM and MM corresponding to a given gene. In the example the probe set is labeled 203508_at (Affymetrix nomenclature) which corresponds to the gene 1B of the tumor necrosis factor receptor superfamily. In Fig. 2(a) we plot the PM and MM intensities versus the probe number. In the specific experiment shown the concentration is 256 pM. Note the PM signals are higher than the MM signals, consistent with thermodynamics expectations. There is a remarkable variability within each set: for instance the PM signals vary from 200 to 2000 (Affymetrix fluorescence units). Fig. 2(b) shows the same data points as in (a) plotted as function of the variable $\Delta G + RT \log \alpha$. The solid line is the Langmuir isotherm corresponding to a concentration of 256 pM. As seen in the figure, the experimental data align quite well along the theoretical prediction. The dotted lines correspond to two Langmuir isotherms at concentrations 64 and 1024 pM, i.e. four times bigger and smaller than the actual experimental concentration. The graph shows the presence of two "outliers" corresponding to data out of the expected trend: these are probes 2 and 5. The target sequence 2 turns out to be highly self-complementary (Fig. 2(c)): a calculation of folding free energies [Heim] shows that the probability of forming a secondary structure is high, which explains probably why the probe 2 has a signal much lower than the predicted Langmuir curve.

Another way of analyzing the experimental data is through a scaling collapse plot. The form given in Eq. (1) suggests that when using a scaling variable $x' = \alpha c \exp(\beta \Delta G)$ all data should collapse onto a single universal master curve of the form $Ax'/(1 + x')$. This is indeed seen in the plots of Fig. 3 which shows some collapse plots for four different probe sets: each plot contains about 300 experimental data points with concentrations ranging from 0.125 to 512 pM. There is a global trend to alignement onto a single curve. Note that only few probes deviate from the theoretical curves as for instance probes 9 and 4 of the probe set 205398_s_at and probe 2 of the probe set AFFX-r2-TagA_at. In Fig. 3 the probe sets 205398_s_at 207655_s_at contain sequences complementary to human genes, while the probe sets AFFX-r2-TagE_at and AFFX-r2-TagA_at contain sequences of artificial origins. The latter spikes were synthesized by Affymetrix and contain sequences which are believed to be unique within the human genome. Our analysis [Heim] shows that the quality of the collapses is the best for sequences of artificial origin. This suggests that the deviations from the Langmuir model have probably biological, rather than thermodynamic origin, i.e. are due to the non-optimal selection of probes.

4. Conclusions. In this paper we reviewed some recent results about the analysis of Affymetrix microarrays data using a physical-chemistry based model [Carlon]. When appropriately rescaled the experimental data tend to collapse to a single universal curve, in agreement with Langmuir thermodynamics. This approach provides an alternative way for determining the gene expression level (i.e. the target concentration c) compared to traditional bioinformatic tools currently available.

The present results were obtained in collaboration with Thomas Heim, Léon-Charles Tranchevent and Gerard Barkema. We acknowledge financial support from the Van Gogh Programme d'Actions Intégrées (PAI) 08505PB of the French Ministry of Foreign Affairs.

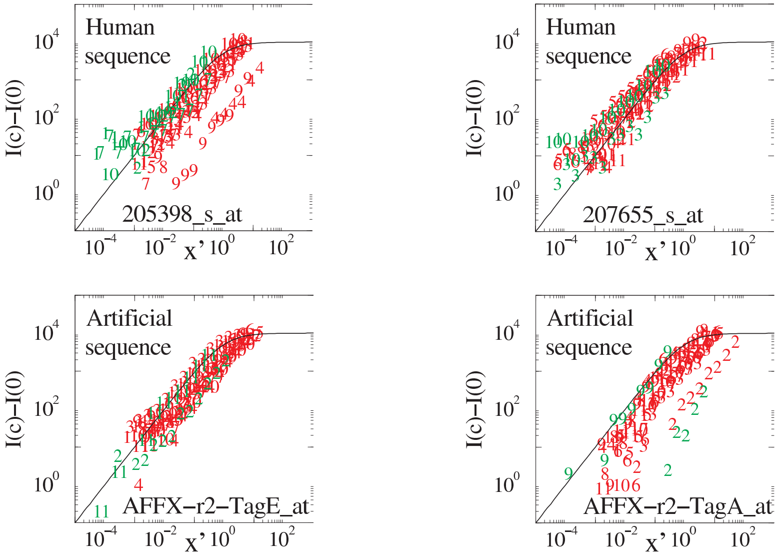

Fig. 3. Collapse plots for 4 different probe sets of the Affymetrix "Latin square" spike-in experiments. The solid line is the Langmuir isotherm (Eq. (1)) plotted as a function of the scaling variable $x' = \alpha c \exp(\beta\Delta G)$.

References

[Carlon] E. Carlon and T. Heim, Physica A 362 (2006), 433.

[Halperin] A. Halperin A. Buhot and E.B. Zhulina, J. Phys.: Cond. Matt. 18 (2006), S463.

[Heim] T. Heim, L.C. Tranchevent, E. Carlon and G.T. Barkema, J. Phys. Chem. B 110 (2006), 22786.

[SpikeLI] D. Baillon, P. Leclercq, S. Ternisien, T. Heim and E. Carlon, SpikeLI: A Bioconductor package for the analysis of Affymetrix spike-in data, http://www.bioconductor.org/packages/1.9/bioc/html/spikeLI.html

[Sugimoto] N. Sugimoto *et al.*, Biochemistry 34 (1995), 11211.

[Xia] T. Xia *et al.*, Biochemistry 37 (1998), 14719.