Buletinul Ştiinţific al Universităţii "Politehnica" din Timişoara

Seria ELECTRONICĂ şi TELECOMUNICAŢII
TRANSACTIONS on ELECTRONICS and COMMUNICATIONS

Tom 50(64), Fascicola 2, 2005

Hybridization modelling in estimation of oligonucleotide microarray expression

Raul Măluţan[1,2] Monica Borda[1] Pedro Gómez Vilda[2] Francisco Díaz[2]

Abstract – Oligonucleotide microarray technology has become widely spread in genetic probing for different purposes. This technology has shown to be highly reliable and dependable, although not free from problems regarding expression estimation, as it relies on the differential or contrasted hybridization of test probes. In many cases hybridization results do not show the expected behavior, thus affecting to the reliability of expression estimation. In the present paper a modeling of the hybridization process is proposed which can be used to predict and correct expression levels on a specific test probe, improving the reliability of the results. Some examples from real microarray processing are shown and discussed.

Keywords: oligonucleotide, hybridization, dynamics

I. INTRODUCTION

DNA microarrays make the use of hybridization properties of nucleic acids to monitor DNA or RNA abundance on a genomic scale in different types of cells. Hybridization process take place between surface-bound DNA sequences, the probes, and the DNA or RNA sequences in solution, the targets. Hybridization is the process of combining complementary, single-stranded nucleic acids into a single molecule. Nucleotides will bind to their complement under normal conditions, so two perfectly complementary strands will bind to each other readily. Conversely, due to the different geometries of the nucleotides, a single inconsistency between the two strands will prevent them from binding.

In oligonucleotide microarrays hundreds of thousands of oligonucleotides are synthesized *in situ* by means of photochemical reaction and mask technology. Probe design in these microarrays is based upon complementarity to the selected gene. An important component in designing an oligonucleotide array is ensuring that each probe binds to its target with high specificity.

The dynamics of the hybridization process underlying genomic expression is complex as thermodynamic factors influencing molecular interaction are still fields of important research and is effects are not taken into account in the estimation of genetic expression by the algorithms currently in use. The present work will present the technical factors that affect expression measurement and are related with hybridization process. Base-pairs interactions, cross-hybridization, kinetic and thermodynamic process are most significant problems relating with microarray assay.

II. NUCLEIC ACIDS

A. *Deoxyribonucleic acid structure*

Deoxyribonucleic acid or DNA can be defined as an antiparallel double helix of nucleotides [8], having deoxyribose as their sugars, linked by sugar-phosphate bonds to adjacent nucleotides in the same chain and by hydrogen bonds to complementary nucleotides in the opposite chain. Also it represents the fundamental substance of which genes are composed. DNA is a nucleic acid that contains the genetic instructions monitoring the biological development of all cellular forms of life, and many viruses. The main role of DNA is to produce protein.
DNA it is organized as two complementary strands [12], head-to-toe, with the hydrogen bonds between them. Each strand of DNA is a chain of chemical "building blocks", called nucleotides, of which there are four types: adenine (A), cytosine (C), guanine (G) and thymine (T). Between the two strands, each base can only bond with one single predetermined other base: A with T, T with A, C with G, and G with C, are the only possible combinations.

Besides DNA there is another nucleic acid called Ribonucleic acid (RNA), also with four bases, but instead of thymine it uses a molecule called uracil (U). Just like T pairs with A, U will pair with A by hydrogen bonding. RNA is used by the cell as temporary copies of portions of DNA. Unlike DNA, RNA is almost always a single-stranded molecule and has a much shorter chain of nucleotides.

[1] Facultatea de Electronică, Telecomunicaţii si Tehnologia Informaţiei, Departamentul Comunicaţii Str. G. Bariţiu Nr. 26-28 Cam. 364, 400027 Cluj-Napoca, e-mail raul.malutan@com.utcluj.ro
[2] Facultad de Informática, Departamento de Arquitectura y Tecnología de Sistemas Informáticos, Campus de Montegancedo s/n, 28660, Boadilla del Monte, Madrid, Spain

An important type of RNA is messenger RNA (mRNA) that encodes and carries information from DNA during transcription to sites of protein synthesis to undergo translation in order to yield a gene product. DNA can be synthesized from a mature mRNA template, and in this case the obtained product it will be a complementary DNA (cDNA). These cDNA sequences are most important components involved in the microarray hybridization process.

B. *DNA hybridization*

Hybridization refers to the annealing of two nucleic acid strands following the base pairing rule. At high temperatures approximately 90ºC to 100°C the complementary strands of DNA separate, denature, yielding single-stranded molecules. Two single strands under appropriate conditions of time and temperature e.g. 65ºC, will renature to form double stranded molecule. Nucleic acids hybrids can be formed between two strands of DNA, two strands of RNA or one strand of DNA and one of RNA. Nucleic acids hybridization is useful in detecting DNA or RNA sequences that are complementary to any isolated nucleic acid.

Finding the location of a gene or gene product by adding specific radioactive or chemically tagged probes for the gene and detecting the location of the radioactivity or chemical on the chromosome or in the cell after hybridization is called in situ hybridization.

In the micrroarray technology [14], hybridization is used in comparing mRNA abundance in two samples, or in one sample and a control. RNA from the sample and control are extracted and labeled with two different fluorescent labels, e.g. a red dye for the RNA from the sample population and green dye for that from the control population. Both extracts are washed over the microarray and gene sequences from the extracts hybridized to their complementary single-strand DNA molecule previously attached to the microarray. Then, to measure the abundance of the hybridized RNA, the array is excited by a laser.

In the oligonucleotide microarrays [13], the hybridization process occurs in the same way, the only difference here is that the sequences to be laid over the chip are sequences of 25 nucleotides length, perfect complementary to same length sequence of the gene, PM – perfect match, and sequences of same length, designed to correspond to PM, but having the middle base, 13th one, changed by its complementary base, MM – mismatch. The MM probes give some estimates of the random hybridization and cross hybridization signals. One thing to be followed in the design of oligonucleotide arrays is ensuring that the probes bind to their target with high accuracy. When the two strands are completely complementary they will bind by a specific hybridization, on contrary if there are mismatches between the nucleotides of the strands and they bind a process called non-specific hybridization or cross-hybridization occurs [7].

Hybridization process has been studied from point of view of interaction between base pairs, the interaction with unintended targets and also from its kinetics processes. Because in practice the DNA chips are immersed in the target solution for a relatively short time, the attainment of equilibrium is not guaranteed. Yet full analysis of the reaction kinetics requires knowledge of the equilibrium state. An understanding of the equilibrium state is also necessary to identify the relative importance of kinetic controls of the performance of the DNA microarrays.

The effect of the cross-hybridization on probe intensity is predictable in the oligonucleotide microarrays, and models for avoiding this have been developed [1], [9], [10], and they will be described further in the paper. Last section of this work will describe a method of modeling the hybridization process based on existing models and unexploited parameters.

III. TECHNICAL FACTORS AFFECTING GENE EXPRESSION MEASUREMENTS

A. *Interaction between pairs*

The nucleic acids duplex stability can be endanger by the interaction between the nucleotide bases. Thermodynamics for double helix formation of DNA/DNA, RNA/RNA or DNA/RNA can be estimated with nearest neighbor parameters. Enthalpy change, ΔH^o, entropy change, ΔS^o, free energy change, ΔG^o, and melting temperature, T_m, were obtained on the basis of the nearest-neighbor model. The nearest-neighbor model for nucleic acids assumes that the stability of a given base pair depends on the identity and orientation of neighboring base pairs [5].

In the nearest-neighbor model, sequence dependent stability is considered in terms of nearest-neighbor doublets. In duplex DNA there are 10 such unique internal nearest-neighbor doublets. Listed in the 5'-3' direction, these are AT/AT TA/TA AA/TT AC/GT CA/TG TC/GA CT/AG CG/CG GC/GC GG/CC. Dimmer duplexes are represented with a slash separating strands in antiparallel orientation e.g., AC/TG means 5'-AC-3' Watson–Crick base-paired with 3'-TG-5'. The ΔG°_{37} can be computed from ΔH^o and ΔS^o parameters using the equation:

$$\Delta G^{o}_{37} = \Delta H^{o} - T\Delta S^{o} \quad (1)$$

As described in [6] the melting temperature T_m is defined as the temperature at which half of the strands are in double helical and half are in the random-coil state. A random-coil state is a polymer conformation where the monomer subunits are oriented randomly while still being bonded to adjacent units.

For self-complementary oligonucleotides, the Tm for individual melting curves was calculated from the fitted parameters using the following equation:

$$T_m = \Delta H^o / \left(\Delta S^o + R \ln C_T\right) \quad (2)$$

Where R is the general gas constant, i.e. 1.987 cal/K mol, the C_T is the total strand concentration, and T_m is given in K. For non-self-complementary molecules, C_T in (2) was replaced by $C_T/4$.

The observed trend in nearest-neighbor stabilities at 37 °C [5] is GC/CG = CG/GC > GG/CC > CA/GT = GT/CA = GA/CT = CT/GA > AA/TT > AT/TA > TA/AT. This trend suggests that both sequence and base composition are important determinants of DNA duplex stability. It has long been recognized that DNA stability depends of the percent G-C content.

B. *Interactions with unintended tags*

As seen in previous sections the major issue in microarray oligonucleotide technology is the selection of probes sequences with high sensitivity and specificity. Because the use of MM probes for assessment of non-specific binding is unreliable [2], a model based on previous mentioned nearest neighbor model has been developed. The positional dependent nearest neighbor model [9] has some modification to the nearest neighbor model, first was to assign different weigh factors at each nucleotide position on a probe with the scope of reflecting that the binding parts of a probe may contribute differently to the stability of bindings, and secondly they took into account two different modes of binding the probes: gene specific binding, *i.e.* formation of DNA-RNA duplexes with exact complementary sequences, and non-specific binding, *i.e.*, formation of duplexes with many mismatches between the probe and the attached RNA molecule.

The model appears to indicate that the two ends of a probe contribute less to binding stability according to their weight factors, and due the mismatch which destabilizes the duplex structure a dip in the gene specific binding weight factors of MM probes around the mismatch position occurs.

C. *Dynamic adsorption model*

The above model, together with the nearest neighbor model solves the problem of binding on microarrays, but still there are factors that affect the gene expression measuring. One of them affects the process of competing adsorption and desorption of target RNA to from probe-target duplexes at the chip surface.

Burden *et al.* [1] develop a dynamic adsorption model based on Langmuir isotherm. If x is the concentration of mRNA target and θ(t) is the fraction of sites occupied by probe-target duplex, then in the forward absorption, target mRNA attaches to probe at a rate $k_f x\left(1-\theta(t)\right)$, proportional to concentration of specific target mRNA and fraction $\left(1-\theta(t)\right)$ of unoccupied probes; and in the backward desroption reaction, target mRNA detaches from probes at a rate $k_b\theta(t)$ proportional to fractions of occupied probes. The fraction of probe sites occupied by probe-target duplexes is then given by the differential equation:

$$\frac{d\theta(t)}{dt} = k_f x\left(1-\theta(t)\right) - k_b\theta(t) \quad (3)$$

For initial condition $\theta(0)=0$, (3) has the following solution:

$$\theta(t) = \frac{x}{x+K}\left[1-e^{-(x+K)k_f t}\right] \quad (4)$$

where $K = k_b/k_f$.

Using (4) Burden *et al.* estimate the measured fluorescence intensity *I*, with I_0 as the background intensity at zero concentration, to be:

$$I(x,t) = I_0 + \frac{bx}{x+K}\left[1-e^{-(x+K)k_f t}\right] \quad (5)$$

At equilibrium, intensity *I*(*x*) at target concentration *x* follows *Langmuir Isotherm* (Fig. 1):

$$I(x) = I_0 + \frac{bx}{x+K} \quad (6)$$

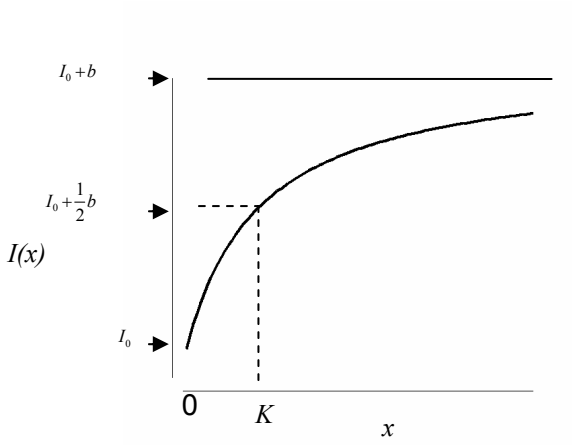

Fig. 1. Hyperbolic response function for the intensity I(x) according to the Langmuir isotherm

IV. HYBRIDIZATION DYNAMICS COMPENSATION

A. *Modeling hybridization by thermodynamics*

It is well known that hybridization processes may be seen under the point of view of general thermodynamic conditions [4], meaning that the hybridization probability of a given test segment will be defined by its thermodynamic conditions, i.e. by its hybridization temperature. Regarding this, one can state that hybridization process will respond to the dynamic equation.

$$P + T \underset{K_b}{\overset{K_f}{\rightleftarrows}} C \quad (7)$$

where P represents the number of oligonucleotides available for hybridization, T the concentration of free RNA target, C the number of bound complexes, k_f and k_b are the respective forward and backwards rate constants for the reaction. This equation has as a natural solution the following expression in the time domain:

$$C(t) = \frac{T}{T+K}\left[1-\exp\left(-t/\tau\right)\right] \quad (8)$$

where K defined as in (4) is an equilibrium dissociation constant, and $\tau = \frac{1}{k_f(T+K)}$ denotes a characteristic time over which the system reaches equilibrium.

Recent studies [3], [15] confirm the hypothesis that hybridization process for the each of the probe pairs follows a time model according to the one from Fig. 2. This model of evolution predicts that the probability of hybridization will be almost zero if not enough time interval is provided for the experiment to take place, and that in the limit, if enough time is allowed saturation will take place.

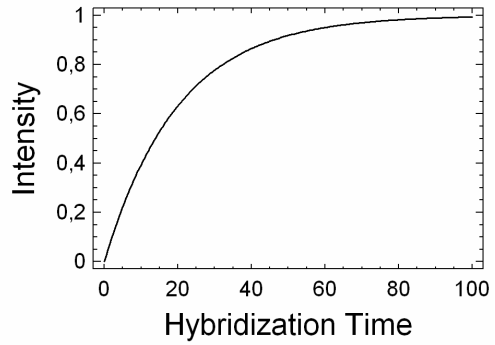

Fig. 2. Theoretical model for perfect match hybridization. Intensity of perfect match versus hybridization time. Figure adapted from [3]

Practical solutions to the different hybridization dynamics could include some of the following procedures:

- Estimate the co-linear and orthogonal components in the PM-MM vectors, and use the co-linear components in estimating the expression.
- Use Independent Component Analysis to correct the PM and MM vectors [11].
- Use multiple regression to convey PM-MM probe pairs to equivalent thermodynamic conditions by processing diachronic hybridization experiments.

The last procedure will be treated more detailed in the following section of the present work.

B. *Exponential regression model*

From (8) we assume that a model to solve the multiple regression problem will have the following form:

$$y = a\left(1-e^{-bx}\right) \quad (9)$$

where a and b are parameters to be estimated adaptively using least square fitting and gradient method.

Vertical least square fitting proceeds by finding the sum of squares of the vertical deviations R^2 of parameters a and b:

$$R^2 = \sum_i \left[y_i - a\left(1-e^{-bx_i}\right)\right]^2 \quad (10)$$

Further let's denote:

$$\varepsilon_i = y_i - a\left(1-e^{-bx_i}\right) \quad (11)$$

With this notation (10) will become

$$R^2 = \sum_i \varepsilon_i^2 \quad (12)$$

The condition of R^2 to be a minimum is that:

$$\begin{aligned} \frac{\partial(R^2)}{\partial a} &= 0 \\ \frac{\partial(R^2)}{\partial b} &= 0 \end{aligned} \quad (13)$$

From (12) and (13) we will get

$$\begin{aligned} \frac{\partial(R^2)}{\partial a} &= \sum_i \varepsilon_i \frac{\partial \varepsilon_i}{\partial a} = -\sum_i \varepsilon_i\left(1-e^{-bx_i}\right) = 0 \\ \frac{\partial(R^2)}{\partial b} &= \sum_i \varepsilon_i \frac{\partial \varepsilon_i}{\partial b} = -\sum_i \varepsilon_i a x_i e^{-bx_i} = 0 \end{aligned} \quad (14)$$

A solution for equations in (14) can be found using the gradient method. In this case the parameters are going to be computed adaptively:

$$a_{k+1} = a_k - \beta_a \frac{\partial(R^2)}{\partial a} = a_k + \beta_a \sum_i \varepsilon_{i,k}\left(1-e^{-b_k x_i}\right) \quad (15)$$

$$b_{k+1} = b_k - \beta_b \frac{\partial(R^2)}{\partial b} = b_k + \beta_b \sum_i \varepsilon_{i,k} a_k e^{-b_k x_i} \quad (16)$$

where $\varepsilon_{i,k}$ is defined as in (11) and β is a parameter used as an adjust step.

V. RESULTS AND DISCUSSION

Having this in mind a data base of hybridization records have been created from experimental data fitted by the before mentioned models. In Fig. 3 the results of hybridization experiments for the same probe test set obtained after two different time intervals have been presented.

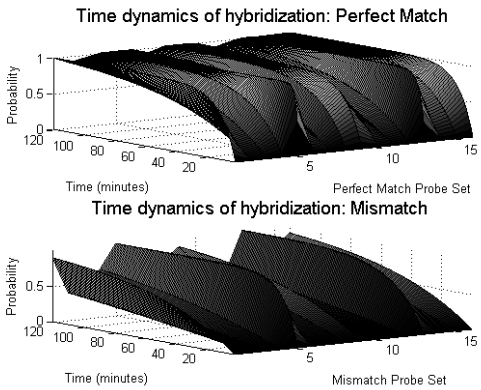

Fig. 3. Time dynamics of hybridization

Fig. 4 shows the regression parameters obtained for time constants. There were extracted the profile of the perfect and mismatch for two different time values so that it can be underline the fact that if enough time is allowed to some probes, the mismatch ones can be corrected

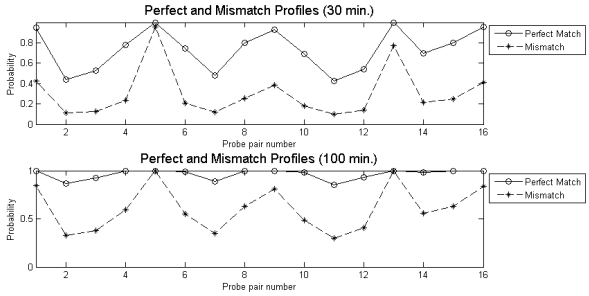

Fig. 4. Perfect and mismatch profiles for time constants. Top template shows the profiles after 30 minutes, and bottom template shows the profile after 100 minutes

REFERENCES

[1] C. Burden, Y.E. Pittelkow, S.R. Wilson, "Statistical Analysis of Adsorption Models for Oligonucleotide Microarrays", *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, article 35, 2004

[2] C. Li, W.H Wong, "Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection", *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 1, January 2001, pp. 31 – 36

[3] H. Dai, M. Meyer, S. Stepaniants, M. Ziman, R. Stoughton, "Use of hybridization kinetics for differentiating specific from non-specific binding to oligonucleotide microarrays", *Nucleic Acids Research*, vol. 30, no. 16, 2002, pp. e86.1 – e86.8

[4] H. El Samad, M. Khammash, L. Petzold, D. Gillespie, "Stochastic Modelling of Gene Regulatory Networks", *Int. Journal of Robust and Nonlinear Control*, vol. 15, issue 15, 2005, pp. 691 – 711

[5] J. SantaLucia Jr, "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics" , *Proc. Natl. Acad. Sci. USA*, *Biochemistry*, vol. 95, February 1998, pp. 1460 – 1465

[6] J. SantaLucia Jr, H.T. Allawi, P.A. Seneviratne, "Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability", *Biochemistry*, vol. 35, no. 11, 1996, pp. 3555 – 3562

[7] J.C. Huang, Q.D. Morris, T.R. Hughes, and B.J. Frey, "GenXHC: a probabilistic generative model for cross-hybridization compensation in high-density genome-wide microarray data", *Bioinformatics*, vol. 21, suppl. 1, 2005, pp. i222 – i231

[8] J.D. Watson, F.H.C. Crick, "A structure of DNA", *Nature*, April 1953, pp. 737

[9] L. Zhang, M.F. Miles, K.D. Aldape, "A model of molecular interactions on short oligonucleotide microarrays", *Nature Biotechnology*, vol. 21, no. 7, July 2003, pp. 818 – 821

[10] N. Sugimoto *et al*., "Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes", *Nucleic Acids Research*, vol. 24, no. 22, 1996, pp. 4501 – 4505

[11] P. Gómez *et al.*, "Robust Preprocessing of Gene Expression Microarrays for Independent Component Analysis", to appear in *Proc. Independent Component Analysis and Blind Signal Separation: 6th International Conference*, 2006

[12] P.P. Vaidyanathan, "Genomics and Proteomics: A signal Processor's Tool", *IEEE Circuits and System Magazine*, Fourth Quarter 2004, pp. 6 – 29

[13] R.J. Lipshutz, S.P.A. Fodor, T.R. Gingeras, D.J. Lockhart, "High density synthetic oligonucleotide arrays", *Nature genetics supplement*, vol. 21, January 1999, pp. 20 – 24

[14] S.K. Moore, "Making chips", *IEEE Spectrum*, March 2001, pp. 54 – 60

[15] Y. Zhang, D.A. Hammer, D.J. Graves, "Competitive Hybridization Kinetics Reveals Unexpected Behavior Patterns", *Biophysical Journal*, vol. 89, 2005, pp. 2950 – 2959