SpikeLI: Analysis of Affymetrix Spike-in data with Langmuir Isotherms

Delphine Baillon, Paul Leclercq, Sarah Ternisien, Thomas Heim and Enrico Carlon

October 17, 2016

Contents

1 Introduction

SpikeLI (pron. spike lee) is a package that performs the analysis of the Affymetrix spike-in data using the Langmuir Isotherm. The aim of this package is to show the advantages of a physical-chemistry based analysis of the Affymetrix microarray data compared to the traditional methods. The spike-in (or Latin square) data for the HGU95 and HGU133 chipsets have been downloaded from the Affymetrix web site in Ref. [**?**]. The model used in the spikeLI package is described in details in Ref. [**?**].

In this section we briefly review the basic Langmuir model and an extension introduced in Ref. [**?**], which takes into account the effect of the hybridization in solution. The latter is the actual model used in the calculation of target concentrations (or expression levels).

1.1 The basic Langmuir model

The Langmuir Isotherm describes the hybridization of target RNA to complementary probes PM or MM. According to this model the fluorescent intensity measured on a feature is given by

$$I = I_0 + \frac{Ace^{\beta\Delta G}}{1 + ce^{\beta\Delta G}} \tag{1}$$

where c is the target concentration, or a measure of the gene expression level that one wants to determine, ΔG is the hybridization free energy, $\beta = 1/RT$ the inverse temperature (R the gas constant), A a scale factor and I_0 the background non-specific signal.

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 Refs. [**?**] and [**?**], using the nearest neighbor model.

In Affymetrix chips the MM probes are generated by changing the nucleotide at the 13th position according to the rules A↔T and C↔G. Therefore in MM probes the mismatches are of the type dA-rA, dT-rU, dC-rC and dG-rG (where d denotes the DNA strand and r the RNA strand). The hybridization experiments in solution [**?**] have shown that the free energies of mismatches depend strongly on the nature and order of the two flanking nucleotides. In total, for the four mismatches given above, there are 64 triplet free energies as there are 16 combinations of flanking nucleotides. Unfortunately only a part of these have been experimentally determined, therefore we had to limit our analysis to those mismatches for which ΔG can be calculated (see Ref.[**?**]).

In Eq. (**??**) we take A and β as fitting parameters. We set $A = 10^4$ and $\beta = 1/RT = 0.74$ mol/kcal corresponding to a temperature of $T = 675$K.

1.2 Including hybridization in solution

A second more refined model (see again for details Ref.[**?**]) is:

$$I = I_0 + \frac{A\alpha c e^{\beta \Delta G}}{1 + \alpha c e^{\beta \Delta G}} \tag{2}$$

which is similiar to the basic langmuir isotherm except for the presence of the factor α, describing the effective reduction of target molecules due to hybridization in solution. The idea is the following: as the target is composed by a complex mixture of sequences there is a possibility that some of the sequences in solution are partially complementary. These will form stable duplexes in solution. Therefore the concentration of target available for hybridization to the array is not c, but αc with $\alpha < 1$ (see Fig. **??**).

For the factor α we use the simple input:

$$\alpha = \frac{1}{1 + \tilde{c} e^{\beta' \Delta G_R}} \tag{3}$$

where $\tilde{c}$ and β' are fitting parameters. The ΔG_R is the free energy of a 25-mer RNA/RNA duplex in solution (data taken from Ref. [**?**]). Note that α is sequence dependent: sequences richer in CG nucleotides will have a stronger affinity for forming stable duplexes in solution.

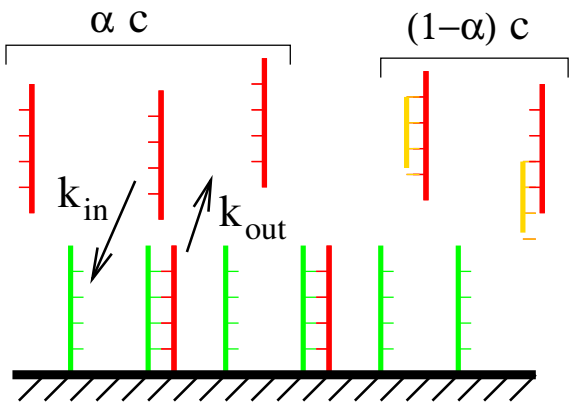

Figure 1: Hybridization model of a set of Microarray probes. The two reactions are 1) Target-probe hybridization and 2) Target-target hybridization in solution leading to an effective reduction of the target concentration to a factor αc.

2 Loading the library and data structure

You install the package using the terminal simply type in:

```
R> CMD INSTALL spikeLI_x.x-x.tar.gz
```

or try to install the package via the menu (for windows and Mac users).

To load the package type

```
R> library("spikeLI")
```

in order to perform some non-linear fits SpikeLI requires the following library as well:

```
R> library("stats")
```

There are three basic functions in this package: collapse(), Ivsc() and IvsDG(). For convenience there are also some vectors giving all probe sets in the HGU95 and HGU133 spike-in sets. These vectors are SPIKE_IN (all spike-in probe sets for the HGU133), SPIKE_INH (human HGU133 spikes), SPIKE_INA (artificial HGU133 spikes), SPIKE_INB (bacterial HGU133 spikes) and SPIKE_IN95. Try the following

```
R> SPIKE_INH
```

```
R> SPIKE_INA
```

The concentration vectors hold the values of the concentration used for the latin square matrix, for the HGU95 and HGU133 Affymetrix Microarrays.

```
R> conc133
```

```
R> conc95
```

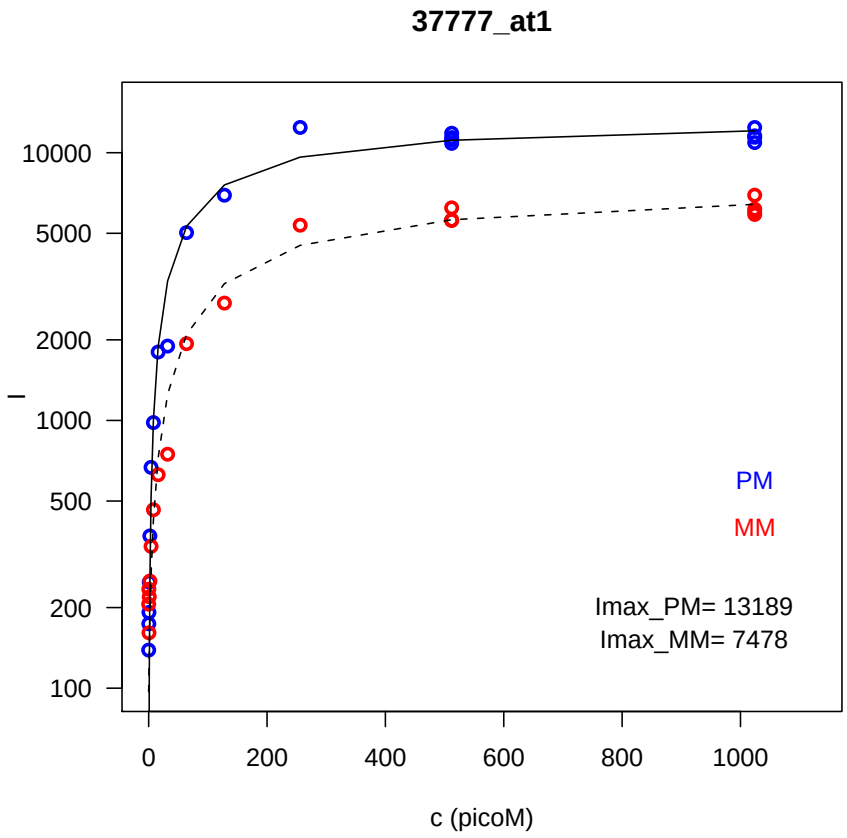

Figure 2: Intensity versus concentration graph for probe 1 of the ”37777” gene

3 Plotting Intensities vs. concentration: Ivsc()

A first function plots the measured intensities as function of the spike-in concentration for a given probe. For instance:

```
R> Ivsc("37777_at",1)
```

shows a plot of intensity vs. spike-in concentration for the probe 1 of the probe set 37777_at (see Fig. **??**). The routine also performs a 3 parameters non-linear fit of the data using the "nls" function of the stats package the formula:

$$I = I_0 + \frac{Ac}{K + c} \tag{4}$$

with I_0, A and K fitting parameters. If the probe number is not given, the routine automatically takes probe=1. If a vector is given for the probe or for the probe set the first elements of both is taken. So, as SPIKE_IN95[1] is "37777_at" typing

```
R> Ivsc(SPIKE_IN95)
```

has the same effects as the above.

The screen displays the fitted value of $I_{max} = I_0 + A$ which is the asymptotic intensity expected for the limit of large concentrations ($c \to \infty$).

4 Plotting Intensities vs. free energies: IvsDG()

This function plots measured intensities at given fixed spike-in concentration as function of the hybridization free energies. The spike-in concentrations range from 0 to 1024 pM (pico molar) for the HGU95 experiment and from 0 to 512 pM for the HGU133. For instance

```
R> IvsDG("AFFX-r2-TagE_at",128)
```

shows two plots (see Fig. **??**) for the probe set "AFFX-r2-TagE_at" at a concentration of 128 pM. On the left side the plot of the PM and MM as functions of the probe number for the given probe set. The data are background subtracted, ie the intensities at zero spike-in concentrations are subtracted from the data. What is plotted in the vertical axis of Fig. **??** is $I(c = 128\text{pM}) - I(0)$. The right side of Fig. **??** shows the same data points plotted as functions of the variable $z = \Delta G - RT \log \alpha$. With this choice the Langmuir isotherm of Eq. (**??**) takes the form:

$$I - I_0 = \frac{Ace^z}{1 + ce^z} \tag{5}$$

which is the curve shown as a solid line (for $c = 128$ pM) in Fig. **??** right. The two dashed lines correspond to concentrations a factor 4 higher and lower than the central one (i.e. 512 and 32 pM). Note that only two MM probes (2 and 11) are shown in Fig. **??**. This is due to the fact, as mentioned above, that the free energies only for a limited numbers of mismatches can be calculated.

All the data in Fig. **??** fall inside the two dashed lines. We identify outliers with this analysis on the basis of their devation from the Langmuir model.

The spike-in concentrations are stored in the two vectors conc95 and conc133. It is possible to generate a simple demo using

```
R> for(i in (2:length(conc133))) IvsDG(SPIKE_IN[1],conc133[i])
```

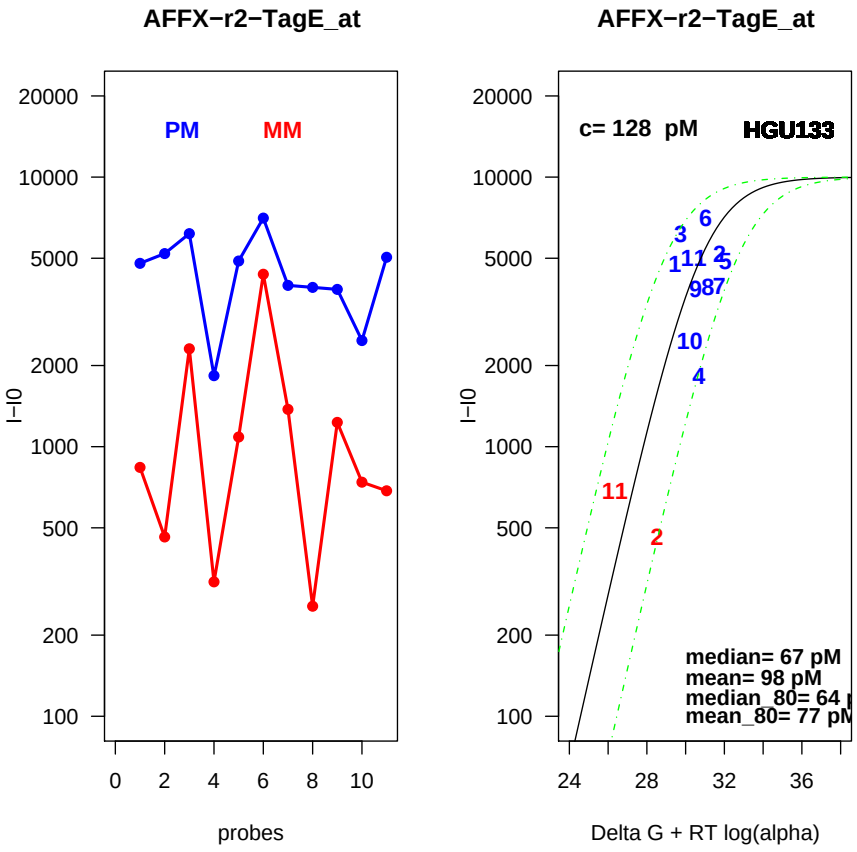

Figure 3: Graph showing the Intensity for each probe and the Intensity versus the Delta G value for each probes for gene "AFFX-r2-TagE"

5 Plotting all concentrations: collapse()

This is the main function of the whole package. It allows to plot on a same graph all spike-in data at different concentrations corresponding to a given probe set. Up to four probe sets can be visualized simultaneously. For instance

```
R> collapse(SPIKE_INA[1:4])
```

displays the collapse plots for the first four probe sets in the vector SPIKE_INA (see Fig. **??**). These are the artificial sequences in the HGU133 experiment: AFFX-r2-TagA_at, AFFX-r2-TagB_at, AFFX-r2-TagC_at and AFFX-r2-TagD_at.

The collapse is simply a plot of the background subtracted intensities $I - I_0$ as a function of the rescaled variable $x' = \alpha c e^{\beta \Delta G}$. When plotted as function of these variables, the data for different concentrations and probes ought to collapse into a single curve which has the form $Ax'/(1 + x')$. The collapse is far from perfect. It works better for some probe sets compared to others. Probes strongly deviating from the Langmuir curve should be considered as outliers (this is for instance probe 2 of the probe set AFFX-r2-TagA_at in Fig. **??**). The plot displays parameters measuring the spreading of the data.

If only one probe set is given in input, as for instance with

```
R> collapse("1091_at")
```

then two graphs are given as output. The graph in the left side is a plot as a function of the variable x', while the one to the right side a plot as a function of the variable $x = ce^{\beta\Delta G}$, which is the collapse plot for the basic Langmuir model (Eq. **??**), in which the effect of hybridization in solution is neglected.

References

[1] Affymetrix download center `http://www.affymetrix.com/analysis/index.affx`

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

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

[4] N. Sugimoto, M. Nakano, and S. Nakano, Biochemistry **39**, 11270 (2000).

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

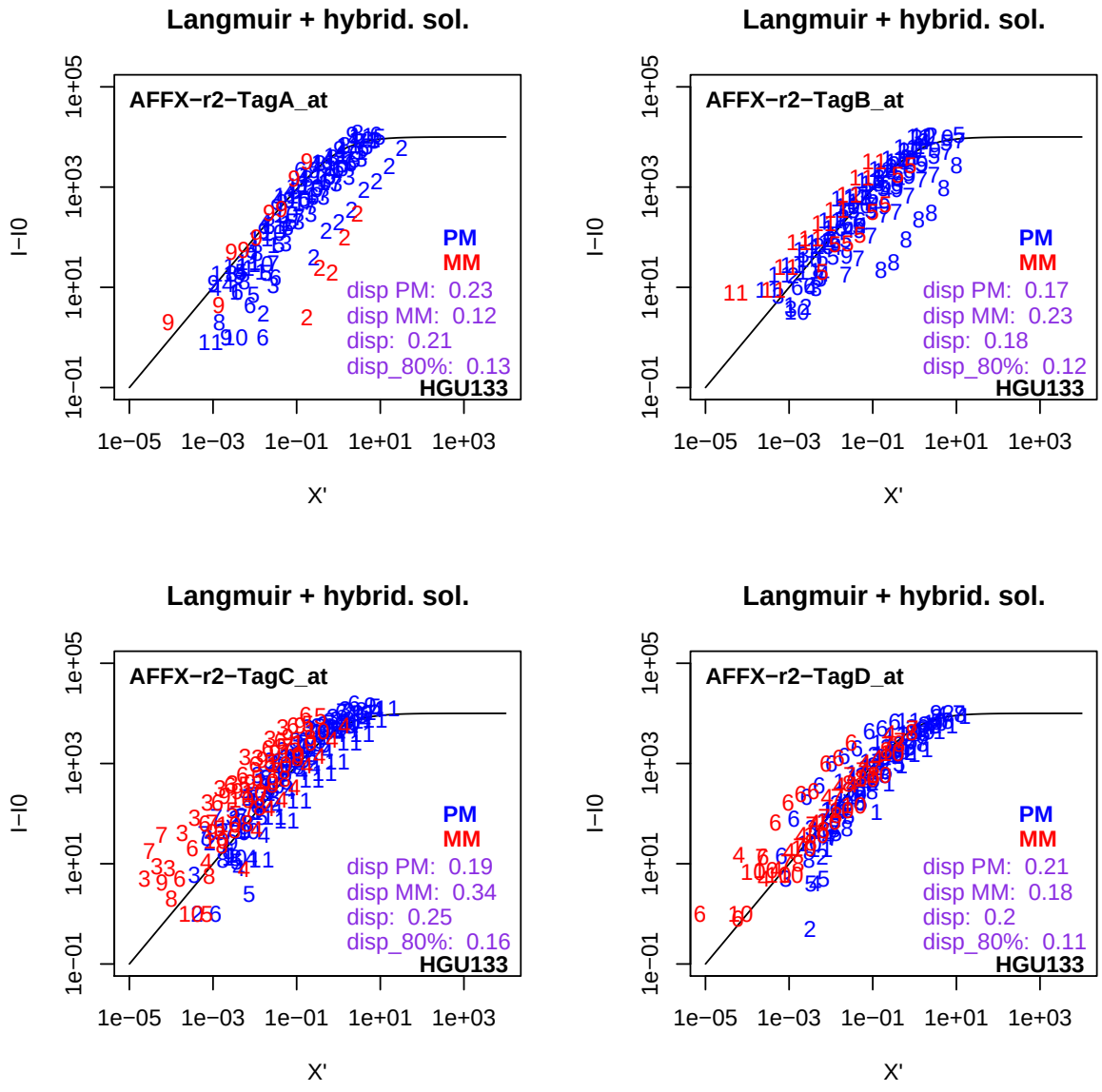

Figure 4: Graph showing the Intensity versus the Hybridization value for each probe on 4 different genes