Modeling and estimation in DNA microarrays

Last time: introduction to DNA microarrays
- massively parallel affinity-based biosensors
- systems which detect/quantify targets of interest after capturing them with probes

The steps preceding a microarray experiment:
(1) array fabrication
   - immobilize probes in unique positions on a surface
(2) sample preparation
   - synthesis of (often labeled) targets from mRNA

Some sources of variations in microarray data:
(1) sample variation
(2) sample preparation
(3) experiment noise (bio-chemical and instrumentation)

Sample variation noise

Essentially, sampling error:

\[
\text{Biological system with } N \text{ analytes, volume } V \\
\rightarrow \text{Biological sample } n \text{ analytes, volume } v
\]

Probability of each analyte being in the sample: \( p_s = v/V \).

Distribution of the number of analytes in the sample:

\[
P(n=k) = \binom{N}{k} p_s^k (1-p_s)^{N-k},
\]

i.e., it has a binomial distribution.
Moments: \( \mu_n = Np_s \), \( \sigma_n^2 = Np_s(1-p_s) \)

If \( \mu_n \) is large, binomial distribution can be approximated by \( \mathcal{N}(\mu_n, \sigma_n^2) \)

Note: sampling uncertainty increases as the sample volume decreases
\[
\frac{\mu_n}{\sigma_n^2} = \frac{\sqrt{\frac{\nu}{V}}}{\sqrt{\frac{1}{n} - \frac{\nu}{V}}}
\]

Sample preparation noise

Labeling (if needed) introduces additional uncertainty

\( p_L \): probability of each analyte being labeled

Then, the number of labeled analytes, \( L \), is distributed as
\[
P(L = k) = \binom{\mu_n}{k} p_L^k (1-p_L)^{\mu_n-k}
\]

Mean and variance: \( \mu_L = \mathbb{E}[L] = n p_L \), \( \sigma_L^2 = np_L(1-p_L) \)

If \( \mu_L \) is large, the above distribution is approx. by \( \mathcal{N}(\mu_L, \sigma_L^2) \)

Experiment (hybridization) noise

Recall the Markov chain model of hybridization process:

\[
\begin{array}{ccc}
\phi & \xrightarrow{P_h} & 1 \\
1-P_r & \rightarrow & P_r \\
\end{array}
\]

state \( \phi \): target molecule binds (hybridizes) to a probe molecule
state \( 1 \): target molecule is free and nearby a probe

\( P_h \): hybridization probability
\( P_r \): probability that a captured molecule is released
Markov chain has a steady-state distribution $\pi$,

$$\pi = \Pi \pi, \quad \Pi = \begin{bmatrix} 1-p_r & p_h \\ p_r & 1-p_h \end{bmatrix}, \quad \pi = \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix}$$

So, the probability that a molecule is captured (once the steady-state has been reached) is $\pi_0$.

Say we have $n_t$ molecules. The number of captured molecules, $n_c$, is distributed as

$$P(n_c = k) = \binom{n_t}{k} \pi_0^k (1-\pi_0)^{n_t-k}$$

Moments:

$$\mu_c = n_t \pi_0 \quad \sigma_c^2 = n_t \pi_0 (1-\pi_0)$$

Recall: if $\mu_c$ large, approximate binomial distr. with normal

$$n_c \sim \mathcal{N}(\mu_c, \sigma_c^2)$$

Measurement model:

$$n_c = \pi_0 n_t + w + \nu,$$

where $w \sim \mathcal{N}(\phi, \phi^2)$ (biochemical noise)

$$\nu \sim \mathcal{N}(\phi, \phi^2)$$ (instrumentation noise)

So, biochemical uncertainty is of shot-noise nature

Note: instead of state “in proximity of a probe”, we could have a state “unbound” but then define

$$P_h = P_{\text{h}} \cdot P_t$$

$P_{\text{h}}$: probability of target-probe hybridization when they are in close proximity

$P_t$: probability that target is in proximity of a probe (usually inversely proportional to the sample volume)
Interference (cross-hybridization)

Consider a microarray with \( N \) different targets, where the number of target \( i \) is \( n_{ti} \).

Denote:
- \( p_i \): the prob. that target \( i \) hybridizes to its probe when nearby
- \( p_{ij} \): the prob. that target \( i \) cross-hybridizes to probe \( j \) when nearby

Markov chain:

In equilibrium:
- \( \mu_i \): steady-state prob. of being in a hybridized state
- \( \mu_{ij} \): steady-state prob. of being in a cross-hybridized state
- \( \mu_{i,k_{i+2}} \): steady-state prob. of being free (unbound)

We can use \( \mu_i \), \( 1 \leq i \leq N \), to find:
\[ q_{li} \]: the prob. target \( i \) is bound to probe \( l \)

In particular:
\[ q_{li} = \begin{cases} 
\mu_{ii} & \text{if target } i \text{ hybridizes with probe } l \\
\mu_{ij} & \text{if } l \text{ is the } j^{th} \text{ probe to which } i^{th} \text{ target cross-hyb.} \\
\phi & \text{otherwise} 
\end{cases} 
\]

\( n_{ti} \): the number of the \( i^{th} \) target molecules captured by the \( l^{th} \) probe

Distribution of \( n_{ti} \):
\[ p(n_{ti} = x) = \binom{n_{ti}}{x} q_{li}^x (1-q_{li})^{n_{ti}-x} \]

As before, approximate with \( N \left( q_{li}; n_{ti}, q_{li} (1-q_{li}) n_{ti} \right) \).
Now, \( n_e = \sum_{i=1}^{N} n_{e_i} \) is the total number of target molecules captured by probe \( l \).

\( n_{e_i} \) - independent \( \Rightarrow n_e \sim \mathcal{N} \left( \sum_{i=1}^{N} g_{ei} n_{t_i}, \sum_{i=1}^{N} g_{ei} (1-g_{ei}) n_{t_i} \right) \).

Define \( \mathbf{m} = [m_1, m_2, \ldots, m_N]^T \). The measurement model:

\[ \mathbf{S} = \mathbf{m} + \mathbf{V}, \quad \mathbf{V} \text{ - instrumentation noise} \]

Equivalently:

\[ \mathbf{S} = \mathbf{Qm}_e + \mathbf{W} + \mathbf{V} \]

\[ \mathbf{Q} = \{g_{ei}\} \text{ - an } N \times N \text{ matrix} \]

\[ \mathbf{W} \sim \mathcal{N} \left( \mathbf{0}, \mathbf{\Sigma}_W \right), \quad \mathbf{\Sigma}_W = \text{diag} \left( \sum_{i=1}^{N} 2_{N_i} (1-2_{N_i}) m_{t_{i}}^2, \ldots, \sum_{i=1}^{N} 2_{N_i} (1-2_{N_i}) m_{t_{N}}^2 \right) \]

\[ \mathbf{V} \sim \mathcal{N} \left( \mathbf{0}, \sigma^2 \mathbf{I} \right) \]