Predicting Microarray Signals by Physical Modeling

Josh Deutsch

University of California

Santa Cruz
Collaborators

Shoudan Liang

NASA Ames

Onuttom Narayan

UCSC
Outline

- Background
  - Gene transcription and regulation
  - What are microarrays
  - Example application: Cancer diagnosis
- The problem: how to improve sensitivity?
- Microarrays in detail
  - Physical constraints on how well they can work
  - Different approaches to signal extraction
- Physical model of the system
  - Different types of binding
  - Experimental comparison
  - Comparison with other approaches
- Conclusions
Gene transcription

DNA → messenger RNA (mRNA)

DNA is read by RNA polymerase

mRNA
Gene transcription

DNA → messenger RNA (mRNA)

- DNA is read by RNA polymerase

- Producing pre-mRNA
Gene transcription

DNA → messenger RNA (mRNA)
- DNA is read by RNA polymerase
- Producing pre-mRNA
- Which is spliced to get rid of junk
Gene transcription

The ribosome translates every three base pairs into one amino acid e.g. CAU → Histidine.
Gene transcription

The ribosome translates every three base pairs into one amino acid e.g. CAU → Histidine.

(Miller, Hamkalo, and Thomas)
Many genes and cell types use the same proteins to regulate transcription. A specific combination of these is probably needed to turn on a gene.
Genetic networks

The state of the cell can be understood from the way different proteins regulate each other and respond to external inputs.
What are microarrays?

Gene 1

Fabricate a chip by synthesizing 25 base pair oligomers attached to a substrate.

Gene 2

Gene 3

Gene 4
What are microarrays?

Fabricate a chip by synthesizing 25 base pair oligomers attached to a substrate.
The oligomer probes are attached to the surface via polymeric linker molecules. Their density is about 1 probe every 40 square angstroms.
Sample preparation

Extract mRNA from cell samples
Sample preparation

Extract mRNA from cell samples

mRNA → reverse transcriptase → cDNA

mRNA
Hybridization to array

Add fluorescent tags to targets:
Hybridization to array

Add fluorescent tags to targets:

Then hybridize to probes:
Linkage to substrate


In many experiments, target molecules can be longer or shorter than probe molecules. When they’re longer, secondary structure formation can also occur.
A real microarray

From Liming Shi /em gene-chips.com
This is a $640 \times 640$ cell U95Av2 chip. Note the vertical bands which are evidence of misalignment with the scanner.
Application: Cancer diagnosis

Microarrays have been used to differentially diagnose different kinds of cancer, for example:

- Diffuse large B-cell Lymphoma
- Ovarian Cancer
- Leukemia
- Breast Cancer
- Small Round Blue Cell Tumors
Application: Cancer diagnosis

Microarrays have been used to differentially diagnose different kinds of cancer, for example

- Diffuse large B-cell Lymphoma
- Ovarian Cancer
- Leukemia
- Breast Cancer
- Small Round Blue Cell Tumors

Often it is hard to distinguish different sub-types of cancer by other means. These distinctions are very important because they often determine the course of treatment.
Recent work by Khan et al has used microarrays to diagnose four types of pediatric tumors, Small Round Blue Cell Tumors (SRBCT):

- neuroblastoma
- Ewing family tumors
- Rhabdomyosarcoma
- non-Hodgkin lymphoma

Using 63 samples for training and 20 for testing, Khan et al were able to correctly predict all data using only 96 genes from a pool of many thousands.
Prediction using minimal gene set

Results for the prediction of samples for different kinds of SRBCT. Using GESSES (J.M. Deutsch Bioinformatics (2003)).
High sensitivity needed

Many genes critical to the function of a cell are only present in minute concentrations, for example in the initial stages of a regulatory cascade.

With many orders of magnitude of concentrations present for different kinds of mRNA, this pushes the limits of microarray technology.
High sensitivity needed

Many genes critical to the function of a cell are only present in minute concentrations, for example in the initial stages of a regulatory cascade.

With many orders of magnitude of concentrations present for different kinds of mRNA, this pushes the limits of microarray technology.

Data from “1532 Series” Affymetrix spike-in human gene experiments.
High sensitivity needed

Many genes critical to the function of a cell are only present in minute concentrations, for example in the initial stages of a regulatory cascade.

With many orders of magnitude of concentrations present for different kinds of mRNA, this pushes the limits of microarray technology.

Data from “1532 Series” Affymetrix spike-in human gene experiments.
Is this variation all noise?

This variation is reproducible:
Is this variation all noise?

This variation is reproducible:
Is this variation all noise?

This variation is reproducible:

![Graph showing reproducible variation](image)

It is likely due to the physicochemical differences between target and probe molecules.

- The output measures the amount of binding.
- In the case of these Affymetrix arrays, the hybridization is between cRNA targets in solution and DNA probes.
- Clearly not all target molecules are binding to the probes.
Why not design it to be less variable?

It is hard finding optimal conditions for a microarray to work. An important parameter is the *melting temperature* for DNA/RNA hybridization of a probe and complementary target molecule.

At temperatures $<<$ melting temperature, the time scale for relaxation becomes very long. The system becomes too sticky and irreversible. (Oligomers are 25 base pairs).
Why not design it to be less variable?

It is hard finding optimal conditions for a microarray to work. An important parameter is the *melting temperature* for DNA/RNA hybridization of a probe and complementary target molecule.

- At temperatures << melting temperature, the time scale for relaxation becomes very long. The system becomes too sticky and irreversible. (Oligomers are 25 base pairs).
- At temperature >> melting temperature, nothing hybridizes.
Why not design it to be less variable?

It is hard finding optimal conditions for a microarray to work. An important parameter is the *melting temperature* for DNA/RNA hybridization of a probe and complementary target molecule.

- At temperatures $<<$ melting temperature, the time scale for relaxation becomes very long. The system becomes too sticky and irreversible. (Oligomers are 25 base pairs).
- At temperature $>>$ melting temperature, nothing hybridizes.
- Therefore it must operate just under a typical melting temperature.
Why not design it to be less variable?

It is hard finding optimal conditions for a microarray to work. An important parameter is the *melting temperature* for DNA/RNA hybridization of a probe and complementary target molecule.

- At temperatures $<<$ melting temperature, the time scale for relaxation becomes very long. The system becomes too sticky and irreversible. (Oligomers are 25 base pairs).
- At temperature $>>$ melting temperature, nothing hybridizes.
- Therefore it must operate just under a typical melting temperature
- The melting temperature depends on the probe’s chemical sequence.
Why not design it to be less variable?

It is hard finding optimal conditions for a microarray to work. An important parameter is the melting temperature for DNA/RNA hybridization of a probe and complementary target molecule.

- At temperatures $<<$ melting temperature, the time scale for relaxation becomes very long. The system becomes too sticky and irreversible. (Oligomers are 25 base pairs).
- At temperature $>>$ melting temperature, nothing hybridizes.
- Therefore it must operate just under a typical melting temperature.
- The melting temperature depends on the probe’s chemical sequence.
- Because of fluctuation in the melting temperature, probes with lower than average melting curves will show less hybridization.
Affymetrix’s approach

mRNA sequence:

...ATTCTCAGGATACTGCCGTATGTCTATGGCTATGGCGCTTCGAATGATACCTCCTCCTACGTATCGATCGGCTTATACGCGATTATACGC...

Probe intensities

Perfect Match

Mismatch

TACTGTCTATGGCTCGCGCTTCGAATG
TACTGTCTATGGACGGCTTCGAATG
Affymetrix uses statistics

- Include probes that have one base altered so that it does not hybridize as readily to the target sequence.
Affymetrix uses statistics

- Include probes that have one base altered so that it does not hybridize as readily to the target sequence.
- Make many probes (typically 16) per mRNA, to average out over this variation as follow:
Affymetrix uses statistics

- Include probes that have one base altered so that it does not hybridize as readily to the target sequence.
- Make many probes (typically 16) per mRNA, to average out over this variation as follow:
  - Affymetrix MAS5.0 uses "One step Tukey’s biweight estimate".
Affymetrix uses statistics

- Include probes that have one base altered so that it does not hybridize as readily to the target sequence.
- Make many probes (typically 16) per mRNA, to average out over this variation as follow:
  - Affymetrix MAS5.0 uses "One step Tukey’s biweight estimate".
  - It subtracts off the mismatch signal from the perfect match in cases where the perfect match is larger.
Affymetrix uses statistics

- Include probes that have one base altered so that it does not hybridize as readily to the target sequence.

- Make many probes (typically 16) per mRNA, to average out over this variation as follow:
  - Affymetrix MAS5.0 uses "One step Tukey’s biweight estimate".
  - It subtracts off the mismatch signal from the perfect match in cases where the perfect match is larger.

- Do statistics on these numbers to try to obtain a good estimate for the input target concentration.
Affymetrix uses statistics

- Include probes that have one base altered so that it does not hybridize as readily to the target sequence.
- Make many probes (typically 16) per mRNA, to average out over this variation as follow:
  - Affymetrix MAS5.0 uses "One step Tukey’s biweight estimate".
  - It subtracts off the mismatch signal from the perfect match in cases where the perfect match is larger.
- Do statistics on these numbers to try to obtain a good estimate for the input target concentration.

This approach ignores the fact that these variations are largely reproducible and depend on the sequence of the probes.
Using sequence information

Zhang, Miles, and Aldape (Nature Biotechnology 2003) invented a model to try to predict hybridization intensities using sequence information.

The model they use postulates an "energy" of binding that depends on position along the sequence. One expects that the ends to be less tightly bound than the middle. There are two kinds of energy:
Zhang, Miles, and Aldape (Nature Biotechnology 2003) invented a model to try to predict hybridization intensities using sequence information.

The model they use postulates an "energy" of binding that depends on position along the sequence. One expects that the ends to be less tightly bound than the middle. There are two kinds of energy:

- Non-specific term that gives the base line binding when the matching target is not present.
Zhang, Miles, and Aldape (Nature Biotechnology 2003) invented a model to try to predict hybridization intensities using sequence information. The model they use postulates an "energy" of binding that depends on position along the sequence. One expects that the ends to be less tightly bound than the middle. There are two kinds of energy:

- Non-specific term that gives the base line binding when the matching target is not present.
- Specific term that has a linear dependence of input target concentration on the measured signal.
Using sequence information

Zhang, Miles, and Aldape (Nature Biotechnology 2003) invented a model to try to predict hybridization intensities using sequence information.

The model they use postulates an "energy" of binding that depends on position along the sequence. One expects that the ends to be less tightly bound than the middle. There are two kinds of energy:

- Non-specific term that gives the base line binding when the matching target is not present.
- Specific term that has a linear dependence of input target concentration on the measured signal.

Minimizing over \((16 \times 2 + 24 \times 2 + 3)\) parameters, they obtain results far better than the MAS5.0 approach of Affymetrix.
Using sequence information

Zhang, Miles, and Aldape (Nature Biotechnology 2003) invented a model to try to predict hybridization intensities using sequence information.

- The model they use postulates an "energy" of binding that depends on position along the sequence. One expects that the ends to be less tightly bound than the middle. There are two kinds of energy:
  - Non-specific term that gives the base line binding when the matching target is not present.
  - Specific term that has a linear dependence of input target concentration on the measured signal.
- Minimizing over \((16\times 2 + 24\times 2 + 3)\) parameters, they obtain results far better than the MAS5.0 approach of Affymetrix.

But does this model capture the physics?
Important physical effects
Important physical effects

Nonspecific Binding
Important physical effects

Target–Target Binding

Nonspecific Binding
Important physical effects

Nonspecific binding and target-target binding are important effects to consider.
Evidence for target target binding

Probes asymptote at different maximum values often correlating with their sequences being similar.
Evidence for nonspecific binding

When the input target concentration is zero or small, there is still a substantial signal from the corresponding probes.
Partial zippering

Because the binding energy is strongly dependent on base pair composition near the melting temperature, we expect that molecules are only partially bound at any one time.
Partial zippering

Because the binding energy is strongly dependent on base pair composition near the melting temperature, we expect that molecules are only partially bound at any one time.

For a segment bound from monomer $n$ to monomer $m$, the free energy of the nearest neighbor model is:

$$
\Delta G_{mn} = \sum_{i=m}^{n-1} \epsilon(i, i + 1) + \epsilon_{initiation}
$$
The partition function for a target molecule hybridized to a probe, \( Z = \exp(-\beta \Delta G) \) gives the total free energy for binding. It sums over all possible zippered states starting at \( m \) and ending at \( m \). For a probe with a total of \( N \) monomers:
Partition function

The partition function for a target molecule hybridized to a probe, $Z = \exp(-\beta \Delta G)$, gives the total free energy for binding. It sums over all possible zippered states starting at $m$ and ending at $m$. For a probe with a total of $N$ monomers:

$$Z = \sum_{m<n=2}^{N} e^{-\beta \Delta G_{mn}} = \sum_{m<n=2}^{N} e^{-\beta (\sum_{i=m}^{n-1} \epsilon(i,i+1) + \epsilon_{\text{initiation}})}$$

sum over all begin and end points
The partition function for a target molecule hybridized to a probe, \( Z = \exp(-\beta \Delta G) \) gives the total free energy for binding. It sums over all possible zippered states starting at \( m \) and ending at \( m \). For a probe with a total of \( N \) monomers:

\[
Z = \sum_{m<n=2}^{N} e^{-\beta \Delta G_{mn}} = \sum_{m<n=2}^{N} e^{-\beta (\sum_{i=m}^{n-1} \epsilon(i,i+1) + \epsilon_{\text{initiation}})}
\]

using the nearest neighbor model.
The partition function for a target molecule hybridized to a probe, $Z = \exp(-\beta \Delta G)$ gives the total free energy for binding. It sums over all possible zippered states starting at $m$ and ending at $m$. For a probe with a total of $N$ monomers:

$$Z = \sum_{m<n=2}^{N} e^{-\beta \Delta G_{mn}} = \sum_{m<n=2}^{N} e^{-\beta(\sum_{i=m}^{n-1} \epsilon(i,i+1) + \epsilon_{\text{initiation}})}$$
Partition function

The partition function for a target molecule hybridized to a probe, $Z = \exp(-\beta \Delta G)$ gives the total free energy for binding. It sums over all possible zippered states starting at $m$ and ending at $m$. For a probe with a total of $N$ monomers:

$$Z = \sum_{m<n=2}^{N} e^{-\beta \Delta G_{mn}} = \sum_{m<n=2}^{N} e^{-\beta \left( \sum_{i=m}^{n-1} \epsilon_{i,i+1} + \epsilon_{\text{initiation}} \right)}$$

This depends strongly on the temperature and the probe sequence.

It can be efficiently computed in $O(N)$ operations using a recursion relation for each probe considered.
Fraction bound

$Z$ (or $\Delta G$) and the chemical potential give the affinity for binding. That is the fraction $f$ of bound probes is

$$f = \frac{1}{1 + e^{\beta(\Delta G - \mu)}}$$

$$\mu = const + \log(\text{concentration}) / \beta$$

is determined by the requirement that

$$\mu_{\text{solution}} = \mu_{\text{probes}}$$
Nonspecific free energy

A probe can still bind, though more weakly to target sequences that have mismatches. The above method of calculation still applies.
Nonspecific free energy

A probe can still bind, though more weakly to target sequences that have mismatches. The above method of calculation still applies.

However we don’t know the concentrations of the target solution because Affymetrix has added uncharacterized concentrations of human RNA as a background.
Nonspecific free energy

A probe can still bind, though more weakly to target sequences that have mismatches. The above method of calculation still applies.

However we don’t know the concentrations of the target solution because Affymetrix has added uncharacterized concentrations of human RNA as a background.

If the binding is weak, when can replace $Z$ with an annealed average over different sequences. This leads to an effective nearest neighbor model for this non-specific binding

$$
\Delta G_{mn}^{NS} = \sum_{i=m}^{n-1} \epsilon_{NS}(i, i + 1) + \epsilon_{NS}^{initiation}
$$

Here the $\epsilon_{NS}$'s have to be empirically determined.
Target-target binding

For similar reasons we don’t know the concentration of targets. Therefore we replace those interactions with an effective annealed model.

\[ \Delta G_{mn}^{TT} = \sum_{i=m}^{n-1} \epsilon^{TT}(i, i + 1) + \epsilon_{initiation}^{TT} \]
The full model

Given an initial set of concentrations for different target molecules, our model will predict the observed intensities of hybridization of targets to probe molecules.

The parameters in the model are:

- Energies in the nearest neighbor model $\epsilon(b_1, b_2)$ e.g. $\epsilon(G, T)$. There are 16 possibilities and 3 sets of terms: specific, non-specific, and target-target.
The full model

Given an initial set of concentrations for different target molecules, our model will predict the observed intensities of hybridization of targets to probe molecules. The parameters in the model are:

- Energies in the nearest neighbor model $\epsilon(b_1, b_2)$ e.g. $\epsilon(G, T)$. There are 16 possibilities and 3 sets of terms: specific, non-specific, and target-target.
- 3 Initiation factors.
The full model

Given an initial set of concentrations for different target molecules, our model will predict the observed intensities of hybridization of targets to probe molecules.

The parameters in the model are:

- Energies in the nearest neighbor model $\epsilon(b_1, b_2)$ e.g. $\epsilon(G, T)$. There are 16 possibilities and 3 sets of terms: specific, non-specific, and target-target.

- 3 Initiation factors.
The full model

Given an initial set of concentrations for different target molecules, our model will predict the observed intensities of hybridization of targets to probe molecules.

The parameters in the model are:

- Energies in the nearest neighbor model $\epsilon(b_1, b_2)$ e.g. $\epsilon(G, T)$. There are 16 possibilities and 3 sets of terms: specific, non-specific, and target-target.

- 3 Initiation factors.

- Small additive background constant,
The full model

Given an initial set of concentrations for different target molecules, our model will predict the observed intensities of hybridization of targets to probe molecules.

The parameters in the model are:

- Energies in the nearest neighbor model $\epsilon(b_1, b_2)$ e.g. $\epsilon(G, T)$. There are 16 possibilities and 3 sets of terms: specific, non-specific, and target-target.

- 3 Initiation factors.

- Small additive background constant,

- Number of probe molecules,
The full model

Given an initial set of concentrations for different target molecules, our model will predict the observed intensities of hybridization of targets to probe molecules. The parameters in the model are:

- Energies in the nearest neighbor model $\epsilon(b_1, b_2)$ e.g. $\epsilon(G, T)$. There are 16 possibilities and 3 sets of terms: specific, non-specific, and target-target.
- 3 Initiation factors.
- Small additive background constant,
- Number of probe molecules,
- Proportionality factor between binding and light intensity.
The full model

Given an initial set of concentrations for different target molecules, our model will predict the observed intensities of hybridization of targets to probe molecules. The parameters in the model are:

- Energies in the nearest neighbor model $\epsilon(b_1, b_2)$ e.g. $\epsilon(G, T)$. There are 16 possibilities and 3 sets of terms: specific, non-specific, and target-target.

- 3 Initiation factors.

- Small additive background constant,
  - Number of probe molecules,
  - Proportionality factor between binding and light intensity.

- A total of 54 parameters, to fit 2464 data points from the Latin Square experiments.
Parameter fitting

We minimize the fitness function:

$$I = \sum_{all\ data} (\log(predicted\ conc.) - \log(\text{observed\ conc.}))^2$$

with respect to all parameters.

This is hard because there are many minima in parameter space. This can be solved using simulating annealing monte carlo, and parallel tempering.
Parallel tempering

Simulate $N$ copies of a system at temperatures $T_1, T_2, \ldots, T_N$. 

Predicting Microarray Signals by Physical Modeling – p.35/39
Parallel tempering

Simulate $N$ copies of a system at temperatures $T_1, T_2, \ldots, T_N$.

Exchange neighboring system configurations depending on the relative energies of the systems $\Delta E$ and the temperature difference $\Delta \beta$ with probability

$$p = \begin{cases} 
\exp(\Delta \beta \Delta E) & \text{for } \Delta \beta \Delta E < 0, \\
1 & \text{otherwise}.
\end{cases}$$
Parallel tempering

- Simulate $N$ copies of a system at temperatures $T_1, T_2, \ldots, T_N$.
- Exchange neighboring system configurations depending on the relative energies of the systems $\Delta E$ and the temperature difference $\Delta \beta$ with probability

$$ p = \begin{cases} 
\exp(\Delta \beta \Delta E) & \text{for } \Delta \beta \Delta E < 0, \\
1 & \text{otherwise}.
\end{cases} $$

It is very efficient at finding low energy states of systems with many degrees of freedom.
Results

$I_{min} = 0.188$
Leave out one

Predict first transcript (16 probes) training on all the other data.
Partial binding

- The ends of the hybridized targets tend to be frayed.
Partial binding

- The ends of the hybridized targets tend to be frayed.
- The fraction of the time a nucleotide is unbound depends on the sequence.
Partial binding

- The ends of the hybridized targets tend to be frayed.
- The fraction of the time a nucleotide is unbound depends on the sequence.
Conclusions

The output of Affymetrix gene chips can be understood in terms of the physicochemical properties of equilibrium statistical mechanics. The processes involve

- Partially zippered RNA targets and DNA probes.
- Non-specific binding of other target molecules.
- Binding of target molecules to each other.
- Taking into account the nonlinearity (saturation) of binding.

This understanding leads a model that predicts mRNA levels well. It is hoped that this model will be turned into software that will benefit researchers wanting a more accurate determination of mRNA levels in their experiments.
Conclusions

The output of Affymetrix gene chips can be understood in terms of the physicochemical properties of equilibrium statistical mechanics. The processes involve

- Partially zippered RNA targets and DNA probes.
Conclusions

- The output of Affymetrix gene chips can be understood in terms of the physicochemical properties of equilibrium statistical mechanics. The processes involve
  - Partially zippered RNA targets and DNA probes.
  - Non-specific binding of other target molecules.
Conclusions

The output of Affymetrix gene chips can be understood in terms of the physicochemical properties of equilibrium statistical mechanics. The processes involve

- Partially zippered RNA targets and DNA probes.
- Non-specific binding of other target molecules.
- Binding of target molecules to each other.
Conclusions

The output of Affymetrix gene chips can be understood in terms of the physicochemical properties of equilibrium statistical mechanics. The processes involve

- Partially zippered RNA targets and DNA probes.
- Non-specific binding of other target molecules.
- Binding of target molecules to each other.
- Taking into account the nonlinearity (saturation) of binding.

It is hoped that this model will be turned into software that will benefit researchers wanting a more accurate determination of mRNA levels in their experiments.
Conclusions

The output of Affymetrix gene chips can be understood in terms of the physicochemical properties of equilibrium statistical mechanics. The processes involve

- Partially zippered RNA targets and DNA probes.
- Non-specific binding of other target molecules.
- Binding of target molecules to each other.
- Taking into account the nonlinearity (saturation) of binding.

This understanding leads a model that predicts mRNA levels well.
Conclusions

- The output of Affymetrix gene chips can be understood in terms of the physicochemical properties of equilibrium statistical mechanics. The processes involve
  - Partially zippered RNA targets and DNA probes.
  - Non-specific binding of other target molecules.
  - Binding of target molecules to each other.
  - Taking into account the nonlinearity (saturation) of binding.

- This understanding leads a model that predicts mRNA levels well.

- It is hoped that this model will be turned into software that will benefit researchers wanting a more accurate determination of mRNA levels in their experiments.