Assessing Effect of Cross-Hybridization on Oligonucleotide Microarrays

S. Kachalo, J.Liang

Dept. of Bioengineering
University of Illinois at Chicago

Abstract

A prediction method to assess non-specific binding based on sequence similarity between probe and target would aid in the understanding and interpreting of global expression profile analysis. In this work we consider a linear hybridization model and estimate the binding coefficients using the quadratic programming technique.

We demonstrate that the estimated binding coefficients are correlated with the similarity of nucleotide sequences between probes and targets. We show that cross-hybridization can be detected for the probes that have 7 or more nucleotide similarity with target.

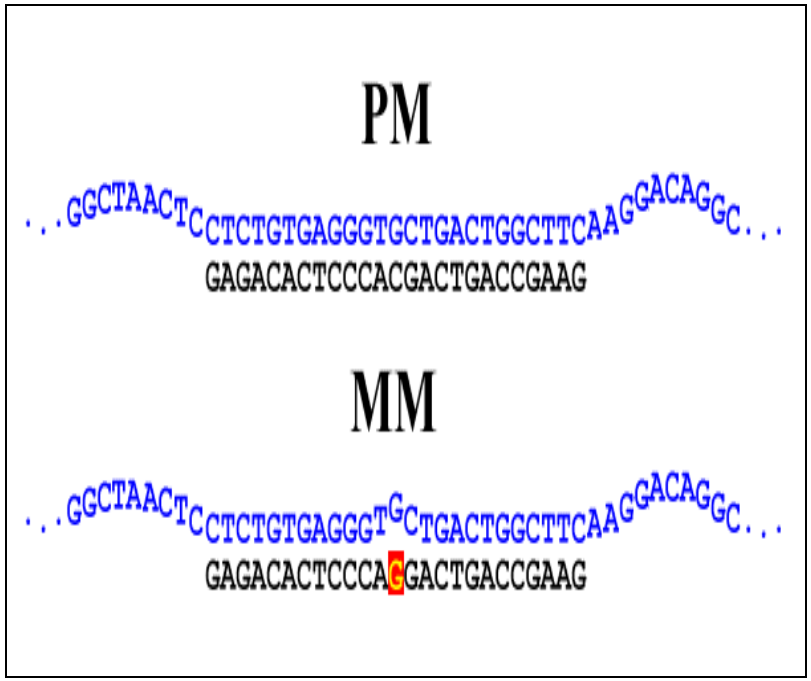
We introduce binding patterns technique for predicting the binding coefficients. Our results suggest that further development based on nucleotide sequence can be fruitful.

Data set

Transcript →		37777_at	684_at	1597_at	38734_at	39058_at	36311_at	36889_at	1024_at	36202_at	36085_at	40322_at	407_at	1091_at	1708_at
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Expts ↓	**A**	0	0.25	0.5	1	2	4	8	16	32	64	128	0	512	1024
	B	0.25	0.5	1	2	4	8	16	32	64	128	256	0.25	1024	0
	C	0.5	1	2	4	8	16	32	64	128	256	512	0.5	0	0.25
	D	1	2	4	8	16	32	64	128	256	512	1024	1	0.25	0.5
	E	2	4	8	16	32	64	128	256	512	1024	0	2	0.5	1
	F	4	8	16	32	64	128	256	512	1024	0	0.25	4	1	2
	G	8	16	32	64	128	256	512	1024	0	0.25	0.5	8	2	4
	H	16	32	64	128	256	512	1024	0	0.25	0.5	1	16	4	8
	I	32	64	128	256	512	1024	0	0.25	0.5	1	2	32	8	16
	J	64	128	256	512	1024	0	0.25	0.5	1	2	4	64	16	32
	K	128	256	512	1024	0	0.25	0.5	1	2	4	8	128	32	64
	L	256	512	1024	0	0.25	0.5	1	2	4	8	16	256	64	128
	M, N, O, P	512	1024	0	0.25	0.5	1	2	4	8	16	32	512	128	256
	Q, R, S, T	1024	0	0.25	0.5	1	2	4	8	16	32	64	1024	256	512

Human portion of Affymetrix Latin Square data set:
59 chips * 409,600 probes;
14 targets with known concentration and unknown complex target in 3 groups of experiments

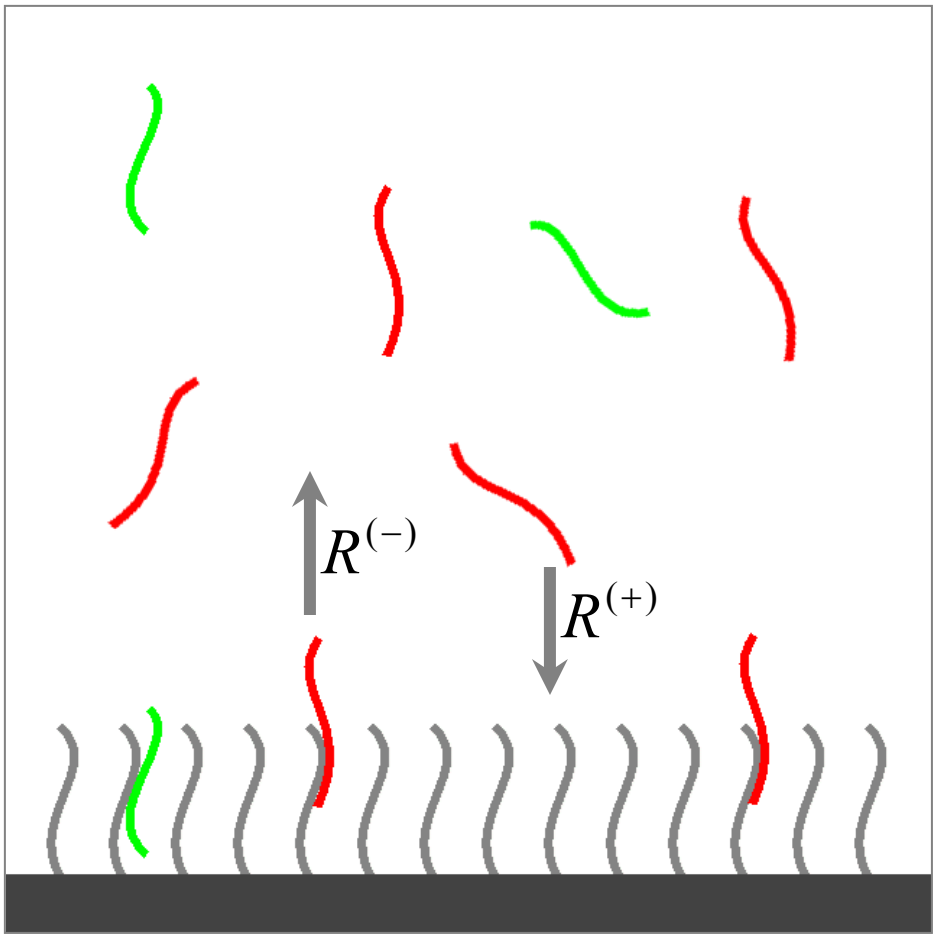
Common assumptions



- Main contribution for PM or MM probe intensity is made by its specific target.
- Non-specific targets binding is about equal for PM and MM probes.

Unfortunately, that is not always true…

DNA binding model



association rate:

$$R^{(+)} \sim N_{unoccupied} X$$

X - target concentration

dissociation rate:

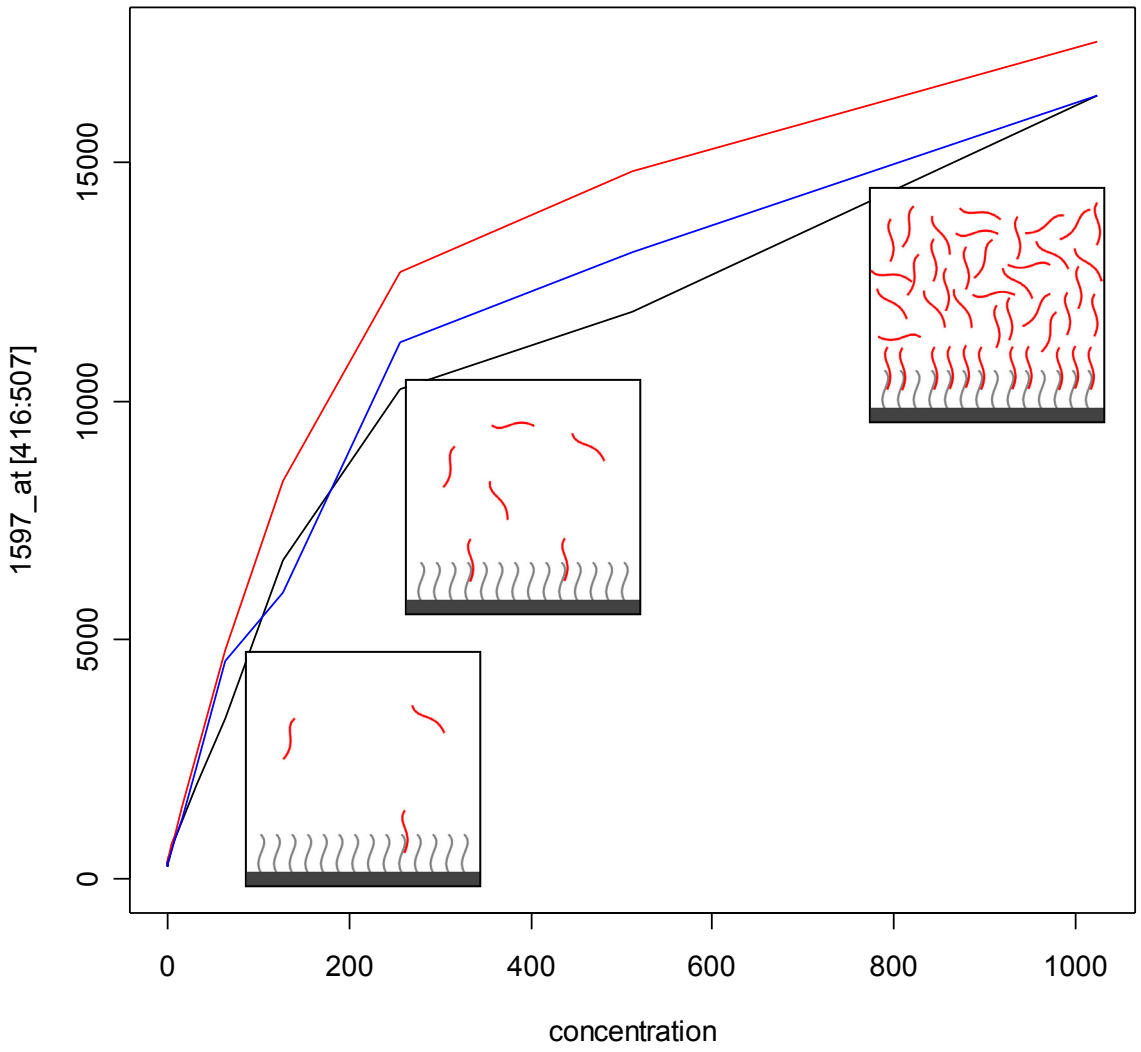
$$R^{(-)} \sim N_{bound}$$

equilibrium:

$$R^{(+)} = R^{(-)}$$

$$N_{bound} \sim N_{unoccupied} X$$

Linear and nonlinear dependency



linear dependency:

$$N_{bound} << N_{unoccupied}$$

$$N_{unoccupied} \approx const$$

$$N_{bound} \sim X$$

saturation:

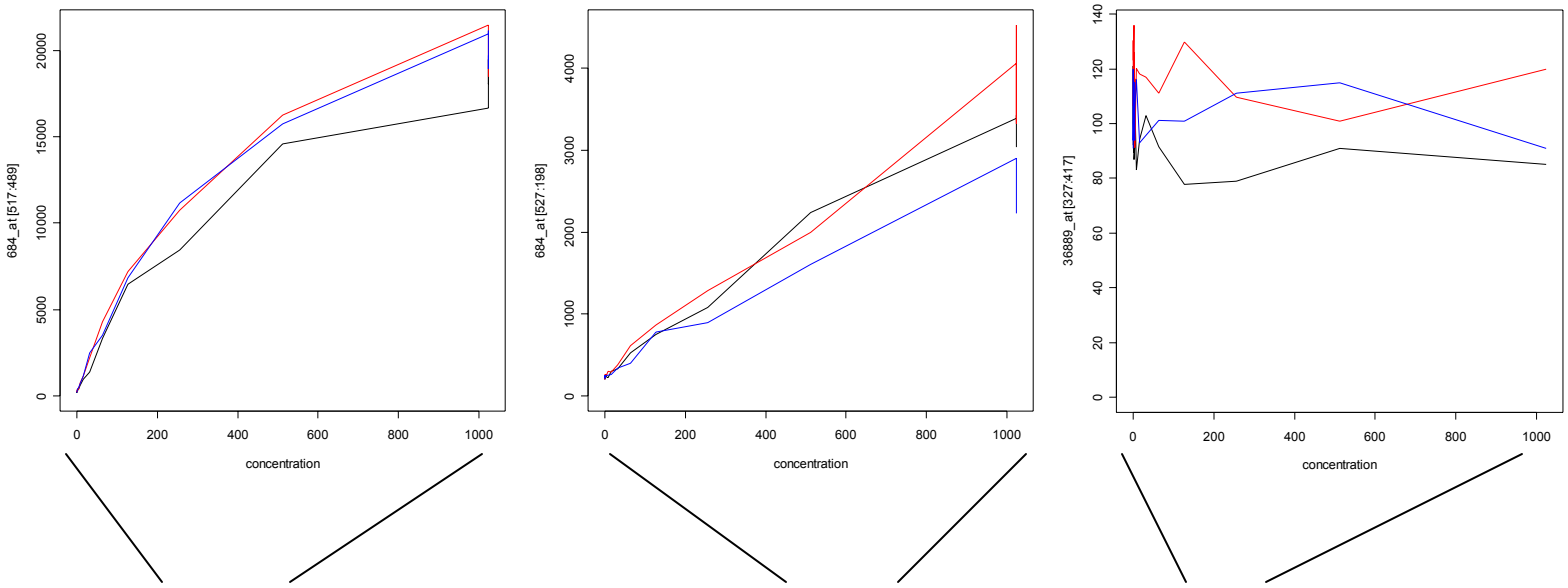
$$N_{bound} >> N_{unoccupied}$$

$$N_{bound} \approx const$$

$$N_{unoccupied} \sim 1/X$$

Dependency is linear if target concentration is low

Distribution of dependencies



PM:	196	11	11	6
MM:	134	28	36	26

Most probes demonstrate typical concentration-intensity curve, linear for low concentrations and nonlinear for higher concentrations

Linear model

$$Y_{ik} = \sum_j B_{ij} X_{jk} + \varepsilon_{ik}$$

$Y_{ik} \geq 0$ - signal intensity of *i*-th probe in *k*-th experiment;

$X_{jk} \geq 0$ - concentration of *j*-th target in *k*-th experiment;

$B_{ij} \geq 0$ - binding coefficient for *i*-th probe and *j*-th target;

ε_{ik} - random noise.

Knowledge of binding coefficients
can reduce calculation of target concentrations
to a simple linear algebra problem!

Calculating binding coefficients

For each probe

$$Y_k = \sum_j B_j X_{jk} + \varepsilon_k$$

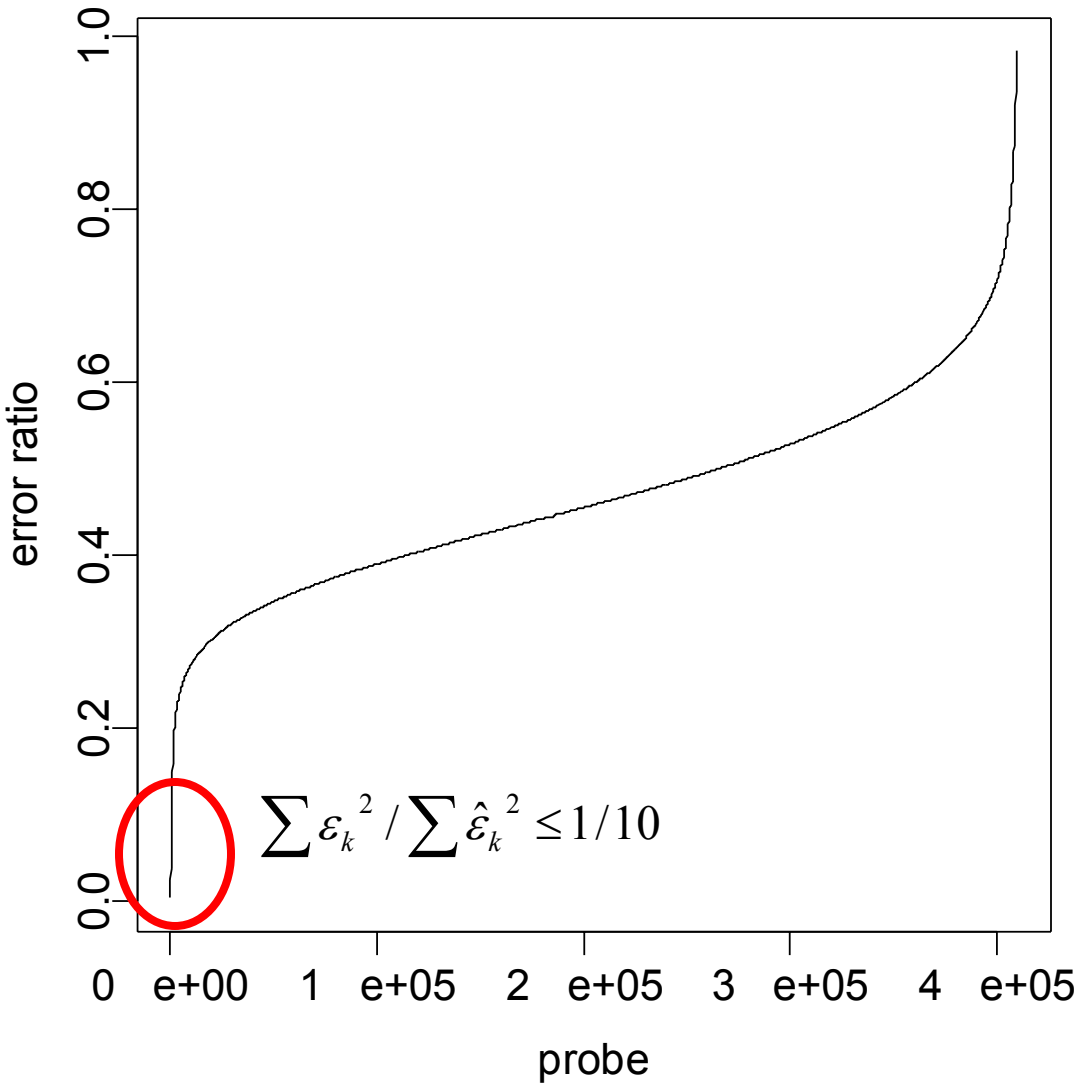
minimize $\sum \varepsilon_k^{\,2}$

subject to $B_j \geq 0$

- it's a quadratic programming problem

random model for comparison:

$$Y_k = \overline{Y} + \hat{\varepsilon}_k$$



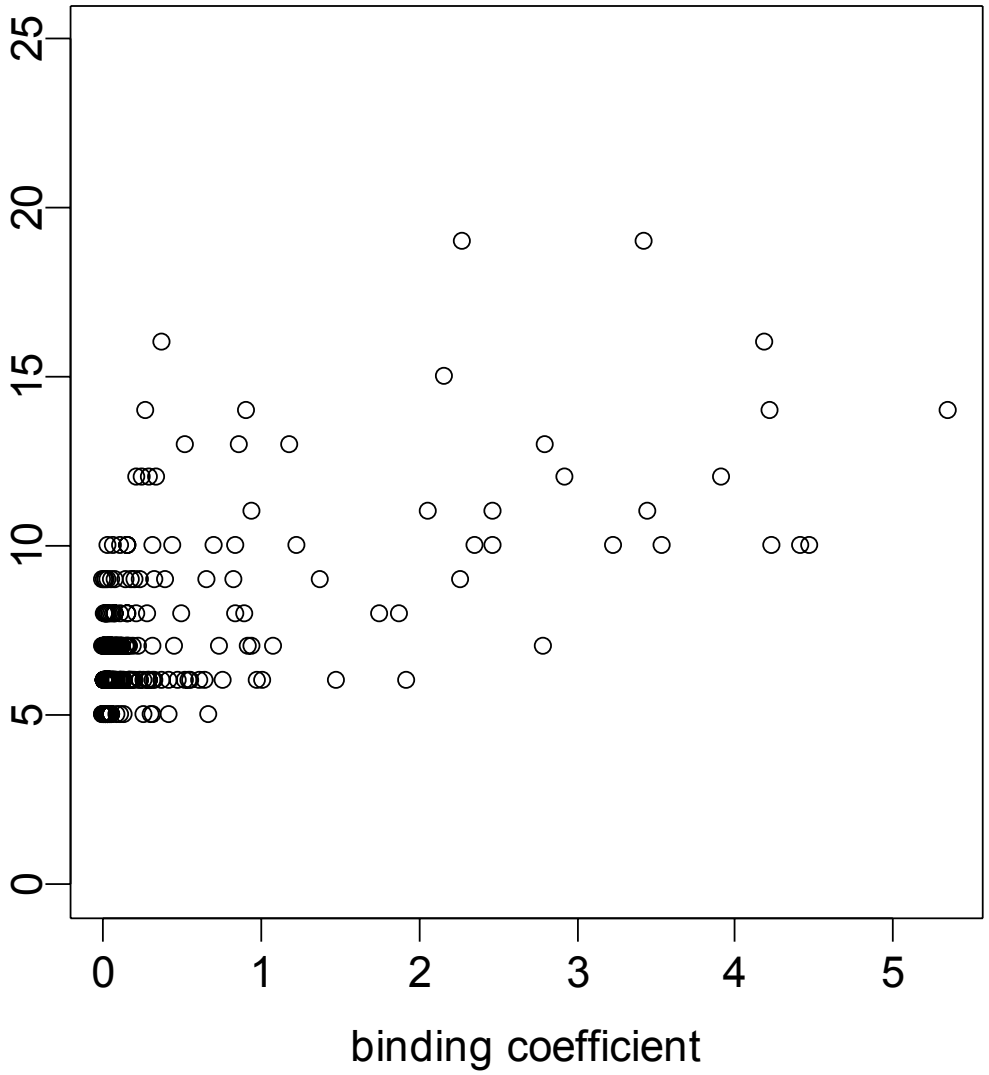
Results

Binding coefficients obtained correlate with sequence similarity measures such as:

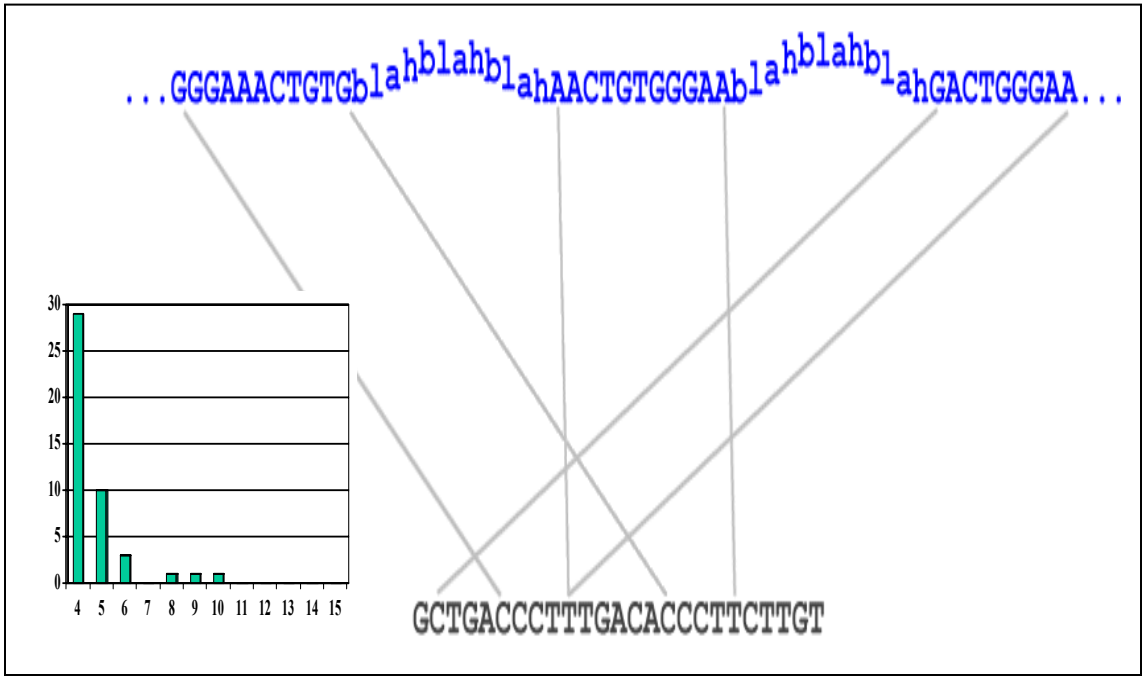
- Longest common substring size

- Smith-Waterman local alignment score

(correlation is over 60%)



Binding patterns contributions



$$B = \sum_a n_a C_a + \varepsilon rror$$

$C_a \geq 0$ - contribution of each type of match into binding coefficient;

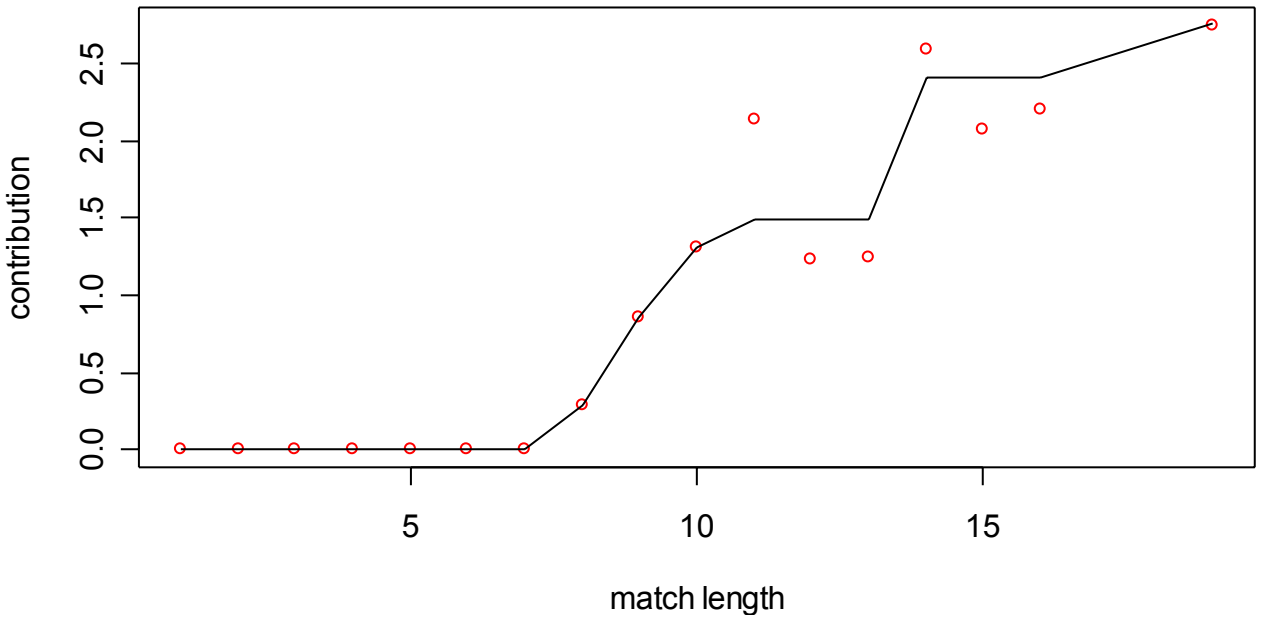
n_a - number of matches of each type.

Calculating contributions

Quadratic programming problem:
minimize total error for all probe-target pairs
under conditions:

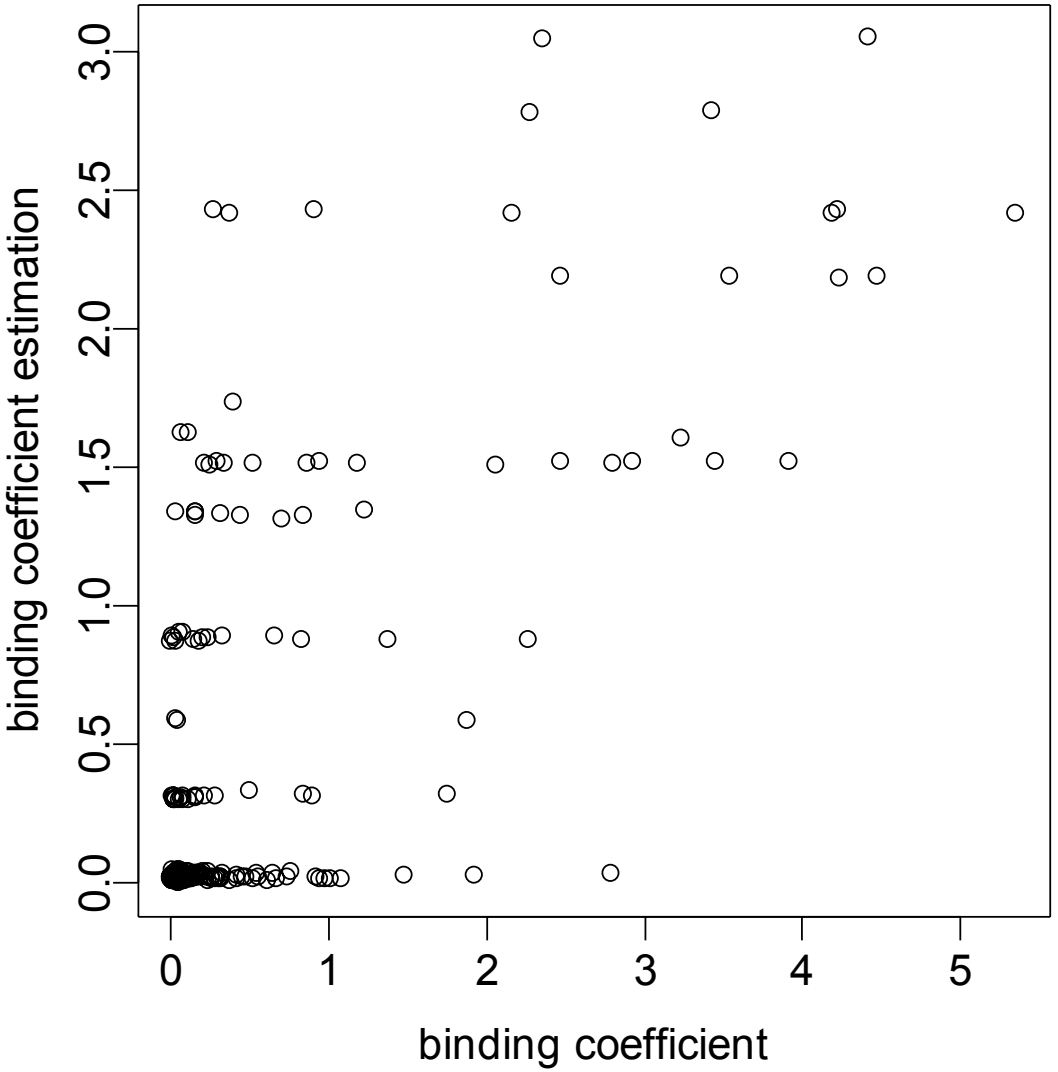
a) $C_L \geq 0$ b) $C_{L+1} \geq C_L$

C_L - contribution of match of length L into binding coefficient;



Estimated binding coefficients

Binding coefficients, estimated via binding patterns contributions are 71% correlated with experimental binding coefficients



Suggestions for Further Experiments

- Lower target concentrations;
- Lower dynamic range of target concentrations;
- Smaller correlation between target concentration - rather random concentrations than ordered Latin Square;
- No complex target.

Summary

- Sequence information should be utilized in microarray data analyses and microarray design;

- Targets with similarities of 7 and more nucleotides to the probe sequence have detectable contribution to its intensity;

- Probe intensity can be assumed linear function of target concentrations for a reasonable range of concentrations;

- If binding coefficients are known, linear binding model can give more accuracy than traditional algorithm.

Thank you!