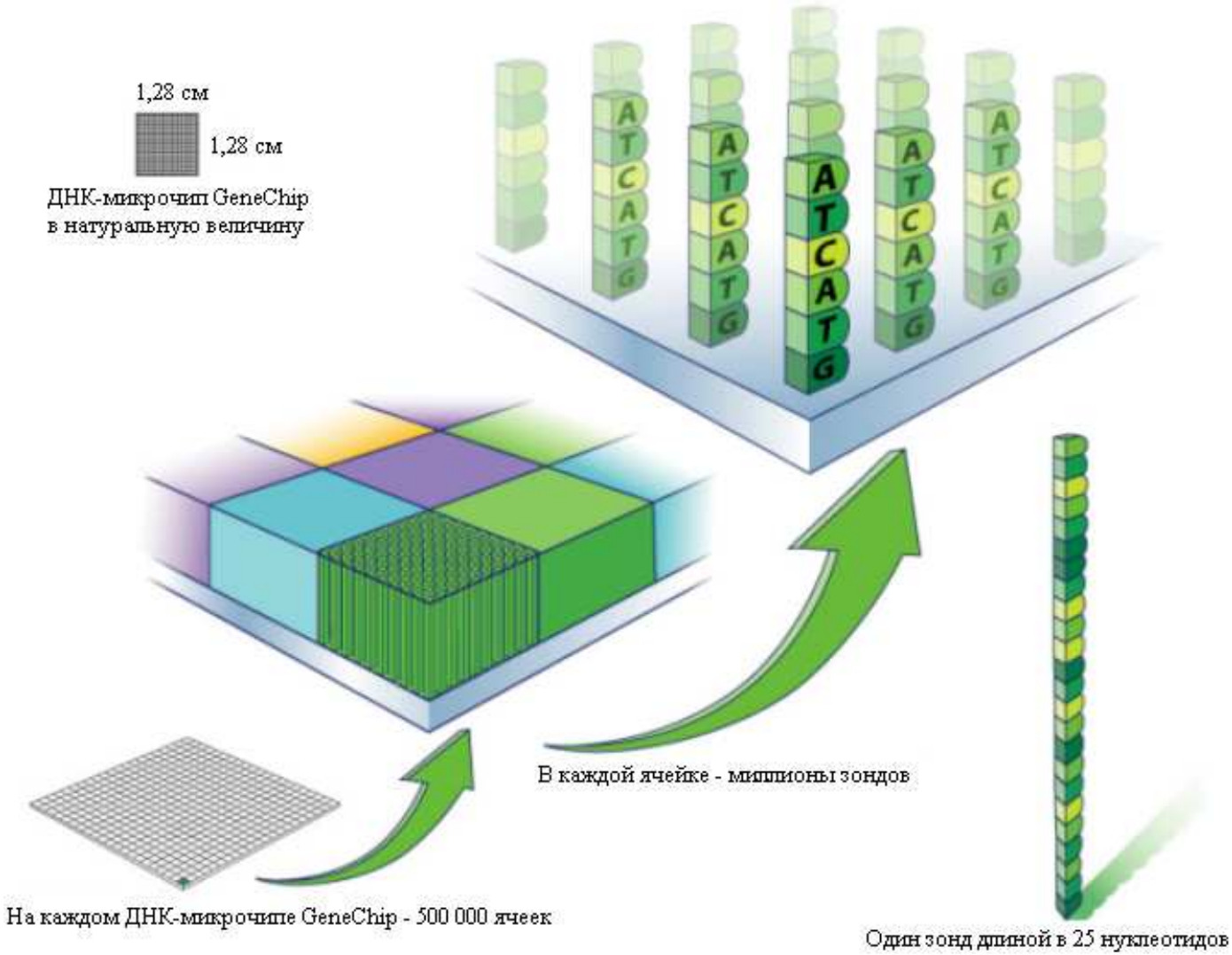
Математическая модель данных микрочипов ДНК и методы оценки её параметров

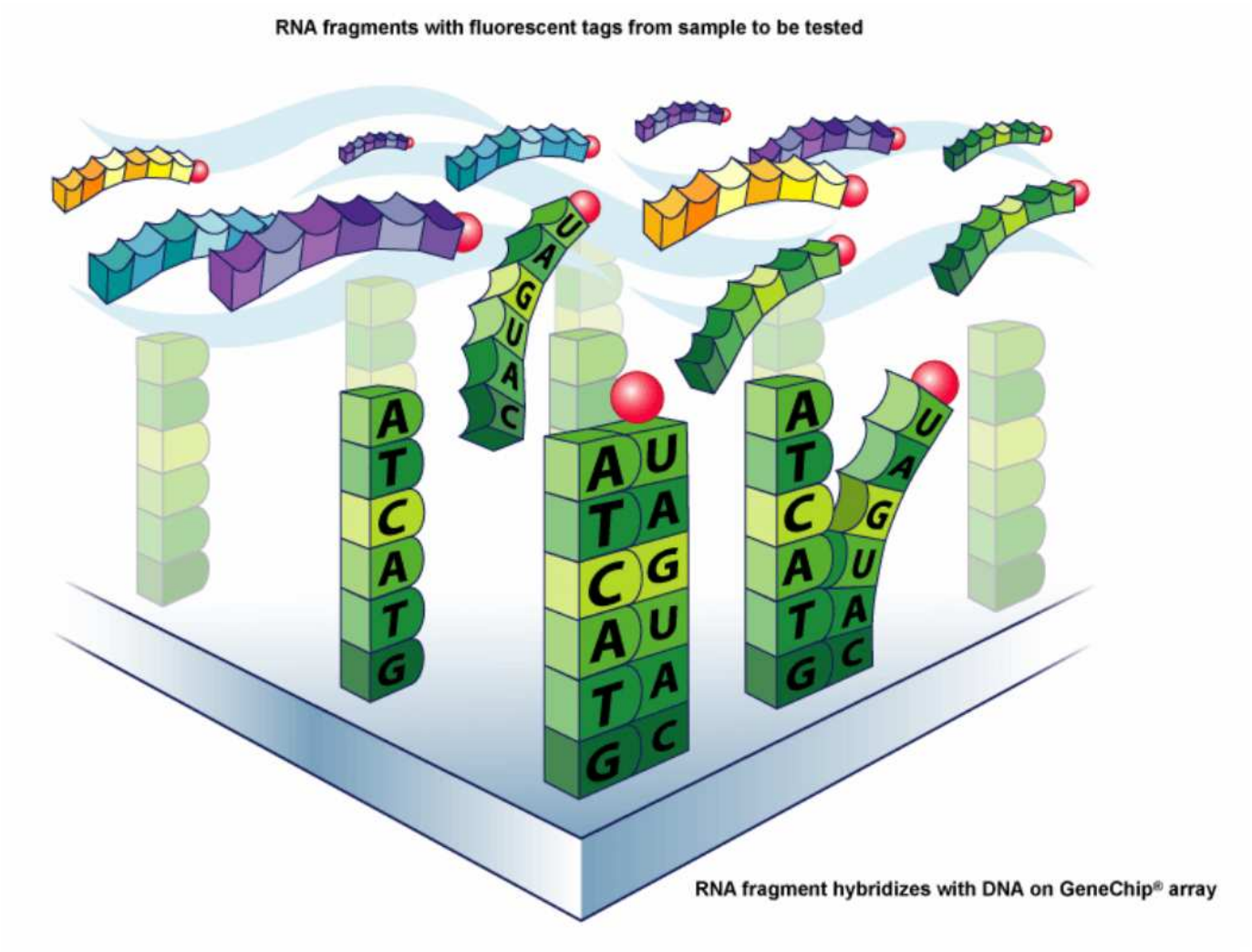
Исполнитель: студентка группы 517 Когадеева М.С.
Научный руководитель: д.ф-м.н. Воронцов К.В.

27 мая 2011

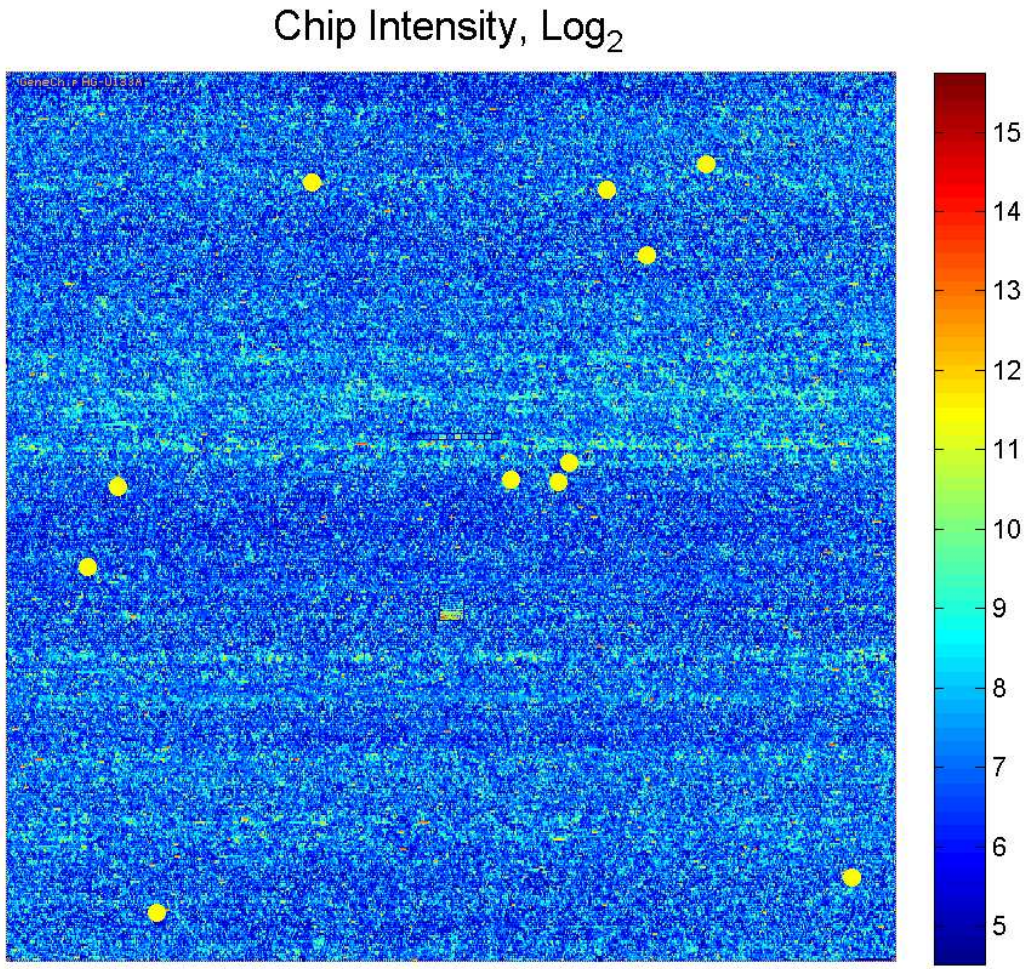
Схема устройства микрочипа класса Affymetrix GeneChip



Процесс гибридизации на чипе



Результат сканирования однокрасочного микрочипа



Задачи анализа данных микрочипов ДНК

Основные задачи:

- По данным об интенсивности свечения проб восстановить исходные концентрации генов в образце
- Сравнить концентрации разных генов на одном чипе
- Сравнить концентрации одинаковых генов на разных чипах

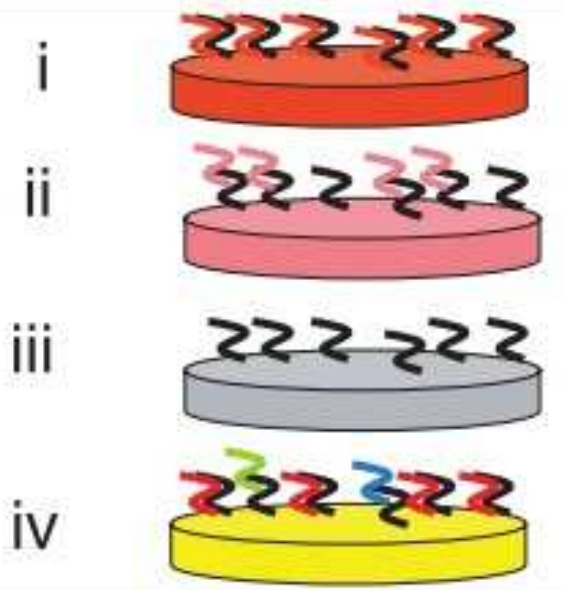
Факторы, которые может учитывать модель интенсивности:

- погрешности измерения, техническая и биологическая вариация;
- неспецифическое взаимодействие — кросс-гибридизация;
- нелинейность зависимости интенсивности от концентрации.

Природа шума:

- техническая вариация — вариация лабораторных условий на этапе выделения материала, его обработки, окрашивания, гибридизации;
- погрешность измерения — неточности работы сканера;
- биологическая вариация — особенности исследуемых организмов и их состояний.

Кросс-гибридизация

Specificity level	Illustration of hybridization
A. probe	
B. spot	

Интенсивность как линейная функция концентрации

Зависимость интенсивности от концентрации генов:

$$I_i^t = d^t \sum_j A_{ij} C_j^t + b_i^t$$

I_i^t — известная интенсивность i-й пробы на t-м чипе,
$C_j^t \geqslant 0$ — концентрация j-го гена на t-м чипе.
$A_{ij} \geqslant 0$ — коэффициент взаимодействия i-й пробы с j-м геном,
d^t — параметр нормализации,
b_i^t — фоновая поправка.

Стандартная задача: $(I, A) \to (C, d, b)$
— восстановление концентраций;

«Сверхзадача»: $(I) \to (A, C, d, b)$
— восстановление концентраций с учётом кросс-гибридизации;

Наша задача: $(I, C) \to (A, d, b)$
— калибровка матрицы A с учётом кросс-гибридизации.

Задача определения коэффициентов A_{ij}

Минимизация квадратичной невязки с L_1-регуляризацией:

$$Q = \sum_t \sum_i \left(I_i^t(A) - Y_i^t\right)^2 \to \min_A,$$

$$\sum_{ij} |A_{ij}| \le \tau,$$

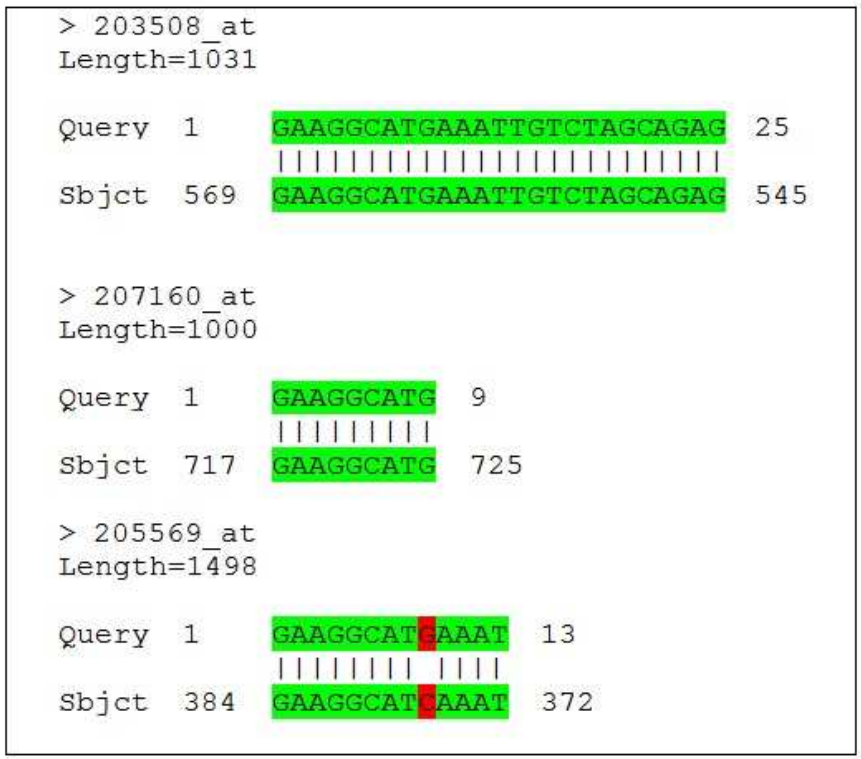
$I_i^t(A)$ — модельная интенсивность i-й пробы на t-м чипе,
Y_i^t — реальная интенсивность i-й пробы на t-м чипе,
$A = (A_{ij})$ — коэффициенты взаимодействия i-й пробы с j-м геном,
τ — параметр регуляризации.

Цель регуляризации: обнулить все коэффициенты, не являющиеся необходимыми для объяснения наблюдаемых интенсивностей.

Предполагается, что коэффициенты, которые останутся ненулевыми, обусловлены только гибридизацией.

Определения коэффициентов A_{ij} путём выравнивания

Матрица A коэффициентов взаимодействия может быть задана априорно из биохимических соображений.
Алгоритм BLASTN оценивает длину участка пересечения и процент схожести участков пробы и гена.



Описание эксперимента

14 образцов, в которых известны концентрации 14 генов.
Для каждого образца приготовлены 3 чипа.
Итого $14 \times 3 = 42$ чипа.

Group ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Gene ID	203508_at 204563_at 204513_s_at	204205_at 204959_at 207655_s_at	204836_at 205291_at 209795_at	207777_s_at 204912_at 205569_at	207160_at 205692_s_at 212827_at	209606_at 205267_at 204417_at	205398_s_at 209734_at 209354_at	206060_s_at 205790_at 200665_s_at	207641_at 207540_s_at 204430_s_at	203471_s_at 204951_at 207968_s_at	AFFX-r2-TagA_at AFFX-r2-TagB_at	AFFX-r2-TagD_at AFFX-r2-TagE_at	AFFX-r2-TagG_at AFFX-r2-TagH_at	AFFX-LysX-3_at AFFX-PheX-3_at
EXP 1	0	0,125	0,25	0,5	1	2	4	8	16	32	64	128	256	512
EXP 2	0,125	0,25	0,5	1	2	4	8	16	32	64	128	256	512	0
EXP 3	0,25	0,5	1	2	4	8	16	32	64	128	256	512	0	0,125
EXP 4	0,5	1	2	4	8	16	32	64	128	256	512	0	0,125	0,25
EXP 5	1	2	4	8	16	32	64	128	256	512	0	0,125	0,25	0,5
EXP 6	2	4	8	16	32	64	128	256	512	0	0,125	0,25	0,5	1
EXP 7	4	8	16	32	64	128	256	512	0	0,125	0,25	0,5	1	2
EXP 8	8	16	32	64	128	256	512	0	0,125	0,25	0,5	1	2	4
EXP 9	16	32	64	128	256	512	0	0,125	0,25	0,5	1	2	4	8
EXP 10	32	64	128	256	512	0	0,125	0,25	0,5	1	2	4	8	16
EXP 11	64	128	256	512	0	0,125	0,25	0,5	1	2	4	8	16	32
EXP 12	128	256	512	0	0,125	0,25	0,5	1	2	4	8	16	32	64
EXP 13	256	512	0	0,125	0,25	0,5	1	2	4	8	16	32	64	128
EXP 14	512	0	0,125	0,25	0,5	1	2	4	8	16	32	64	128	256

Известны концентрации C_j^t и интенсивности I_i^t.

Требуется восстановить коэффициенты взаимодействия A_{ij}.

Качество и непротиворечивость модели

Критерии качества модели:

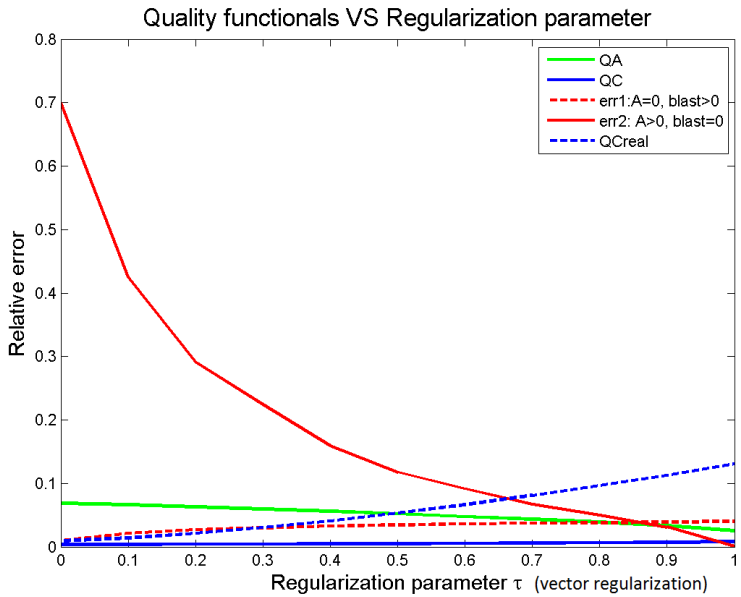
QC(real) — относительная ошибка определения концентрации,

QA(blast) — относительное расхождение с матрицей BLASTN.

Критерии непротиворечивости по техническим репликатам:

QA — относительное расхождение коэффициентов моделей,

QC — относительное расхождение концентраций.



Эффект насыщения в модели интенсивности проб

Зависимость интенсивности от концентрации генов с учетом насыщения:

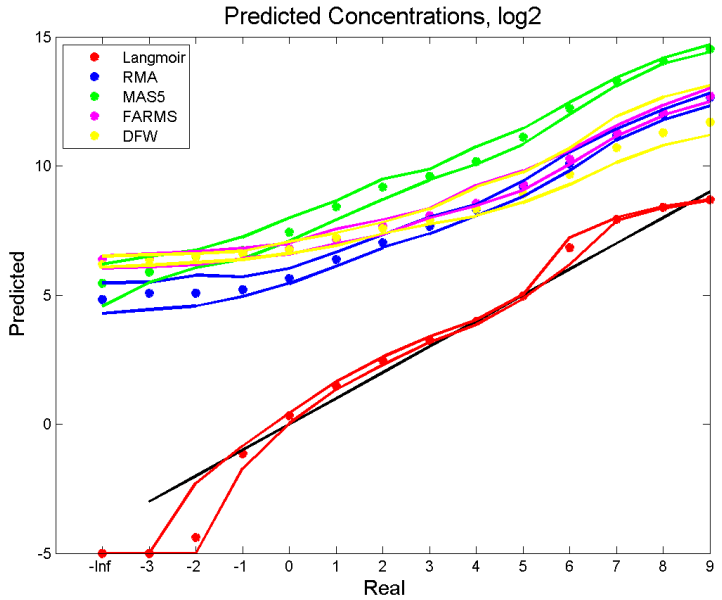
$$I_i^t = d^t \left(\frac{\alpha_i C_{i_0}^t}{1 + \beta_i C_{i_0}^t} + \sum_{j, j \neq i_0} A_{ij} C_j^t + \gamma_i \right) + b_i^t$$

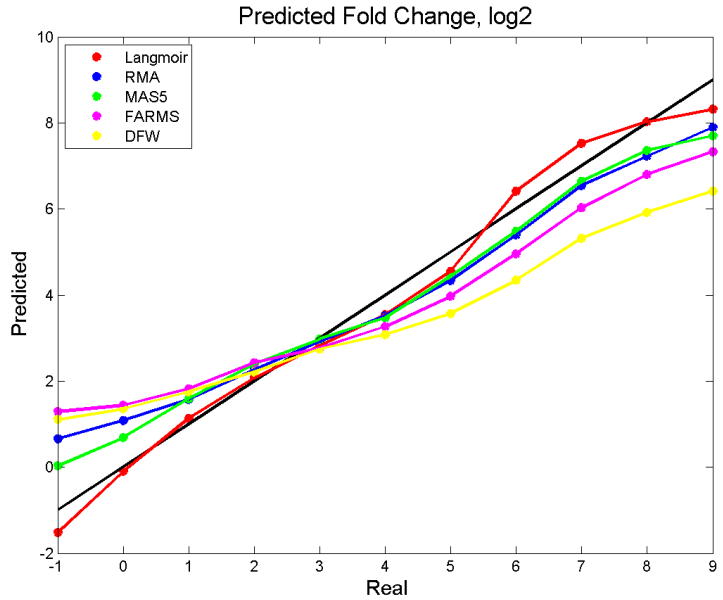
матрица A восстановлена в ходе минимизации квадратичной невязки с ограничениями, накладываемыми алгоритмом выравнивания BLASTN,
α_i, β_i — параметры насыщения,
γ_i — параметр неучтённой кросс-гибридизации,
i_0 — номер специфического гена для пробы i.

Наша задача: $(I, A, d, b) \to (C)$
— восстановление матрицы концентраций с помощью построенной матрицы A с учётом насыщения и кросс-гибридизации.

Сравнение с существующими алгоритмами

MAS5, DFCM: взвешенное среднее интенсивностей соответствующих гену проб
RMA: медиана интенсивностей соответствующих гену проб
FARMS: оценка параметров нормального распределения
Комбинированная модель: использование матрицы взаимодействий A и учёт нелинейной зависимости





Результаты, выносимые на защиту

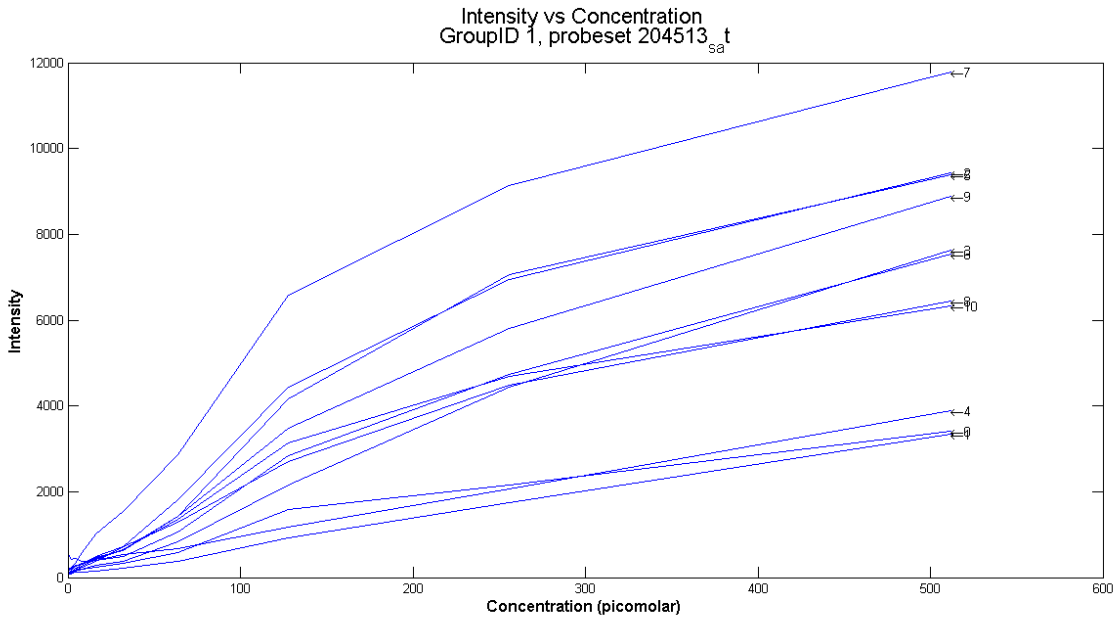
- Предложена модель данных микрочипов ДНК, учитывающая эффекты кросс-гибридизации и нелинейный характер зависимости интенсивности от концентрации генов
- Исследовано влияние регуляризации на точность идентификации параметров модели
- Экспериментально показано, что предложенная модель позволяет восстанавливать относительные концентрации генов не хуже, а абсолютные значения – существенно лучше стандартных методов

Эффект насыщения

Наиболее распространённая модель - модель Ленгмюра

$$I = \frac{\alpha C}{K + C} + b,$$

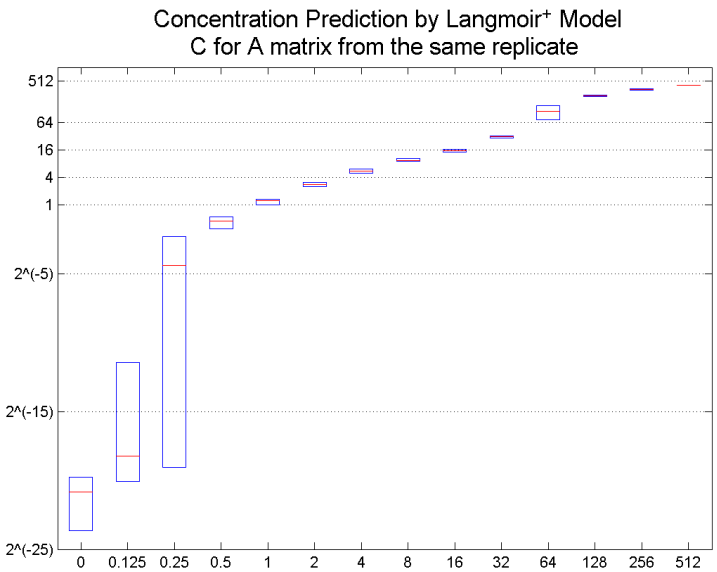
где α и K — параметры насыщения, b — фоновая поправка.

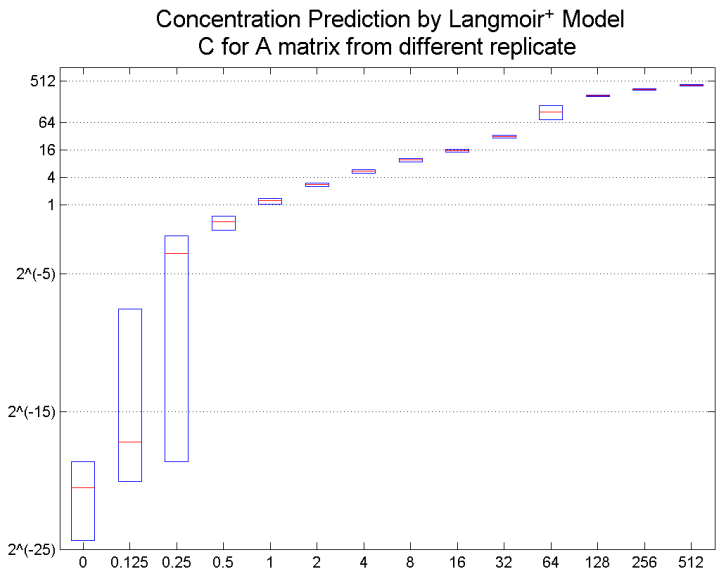


Критерии качества модели

Восстановление концентраций по матрицам взаимодействий A, полученных для разных репликатов путем минимизации функционала

$$Q = \sum_t \sum_i \left(I_i^t(C) - Y_i^t\right)^2 \to \min_C, \qquad C \geq 0$$





Функционалы качества

Критерии качества модели:

QC(real)

$$QC^{real} = \frac{1}{3}\sum_t \frac{\sum_{ij}(\tilde{C}_{ij} - C^t_{ij})}{\sum_{ij}(\tilde{C}_{ij} + C^t_{ij})}$$

— относительная ошибка определения концентрации,где $t_1 < t_2$ - номер репликата, $t_1, t_2 \in \{1, 2, 3\}$.
$\tilde{C}$ — матрица истинных концентраций.

QA(blast)

$$QA^{blast}_{II} = \frac{1}{3}\sum_t \frac{\sum_{ij}[A^t_{ij} > 0][A^{blast}_{ij} = 0]}{|A^{blast}|}.$$

— относительное расхождение с матрицей BLASTN.

Критерии непротиворечивости по техническим репликатам

QA

$$QA = \frac{1}{3}\sum_{t_1,t_2}\frac{\sum_{ij}(A_{ij}^{t_1} - A_{ij}^{t^2})}{\sum_{ij}(A_{ij}^{t_1} + A_{ij}^{t^2})}$$

— относительное расхождение коэффициентов моделей, где $t_1 < t_2$ - номер репликата, $t_1, t_2 \in \{1, 2, 3\}$.

QC

$$QC = \frac{1}{3}\sum_{t_1,t_2}\frac{\sum_{ij}(C_{ij}^{t_1} - C_{ij}^{t_2})}{\sum_{ij}(C_{ij}^{t_1} + C_{ij}^{t_2})}$$

— относительное расхождение концентраций.