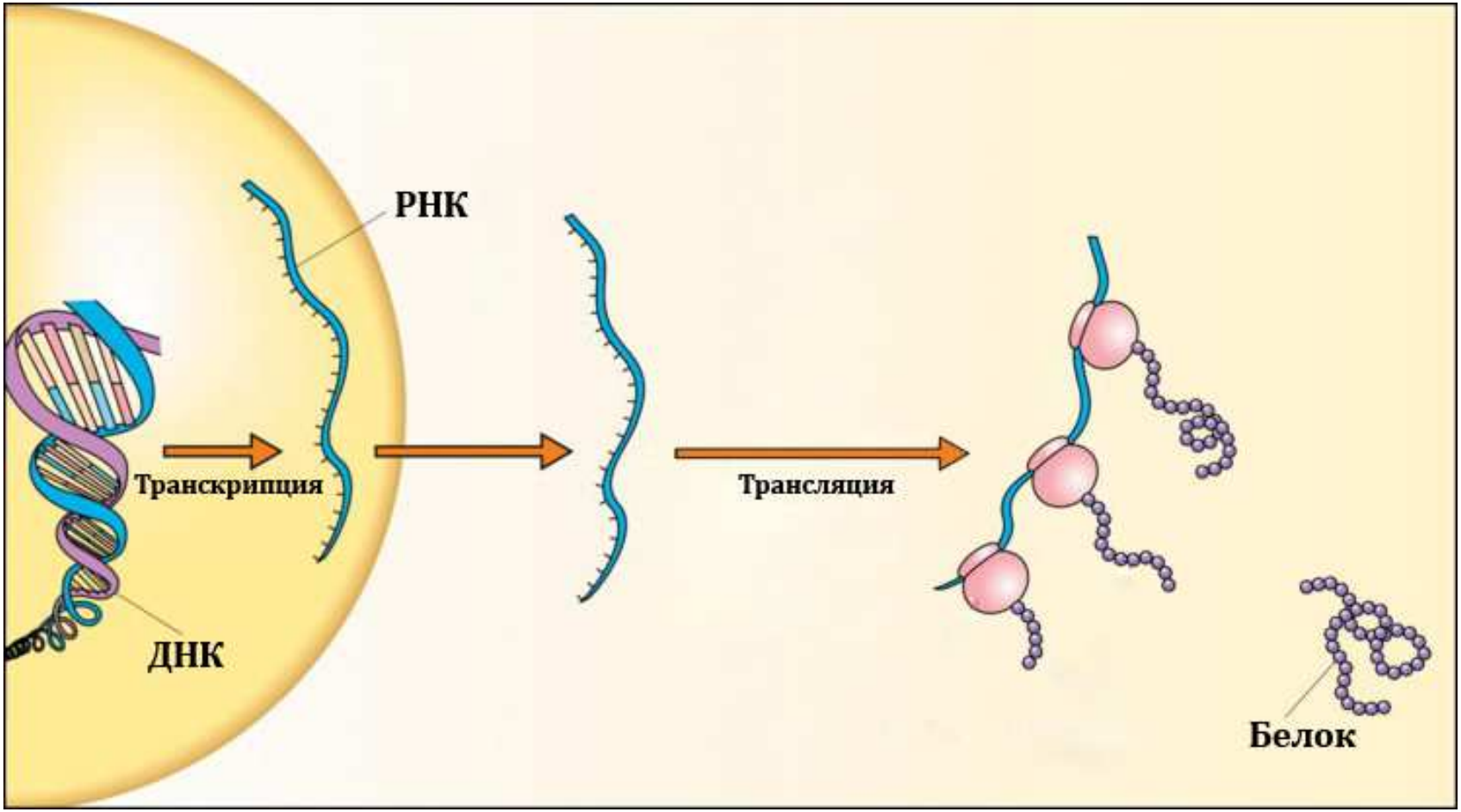
Оптимизационные методы решения некорректно поставленных задач анализа данных ДНК-микрочипов

Рябенко Евгений, аспирант ВМК МГУ
научный руководитель д.ф.-м.н., К.В. Воронцов

Семинар отдела Интеллектуальных систем
ВЦ РАН • 12 сентября 2012 г.

Центральная догма молекулярной биологии

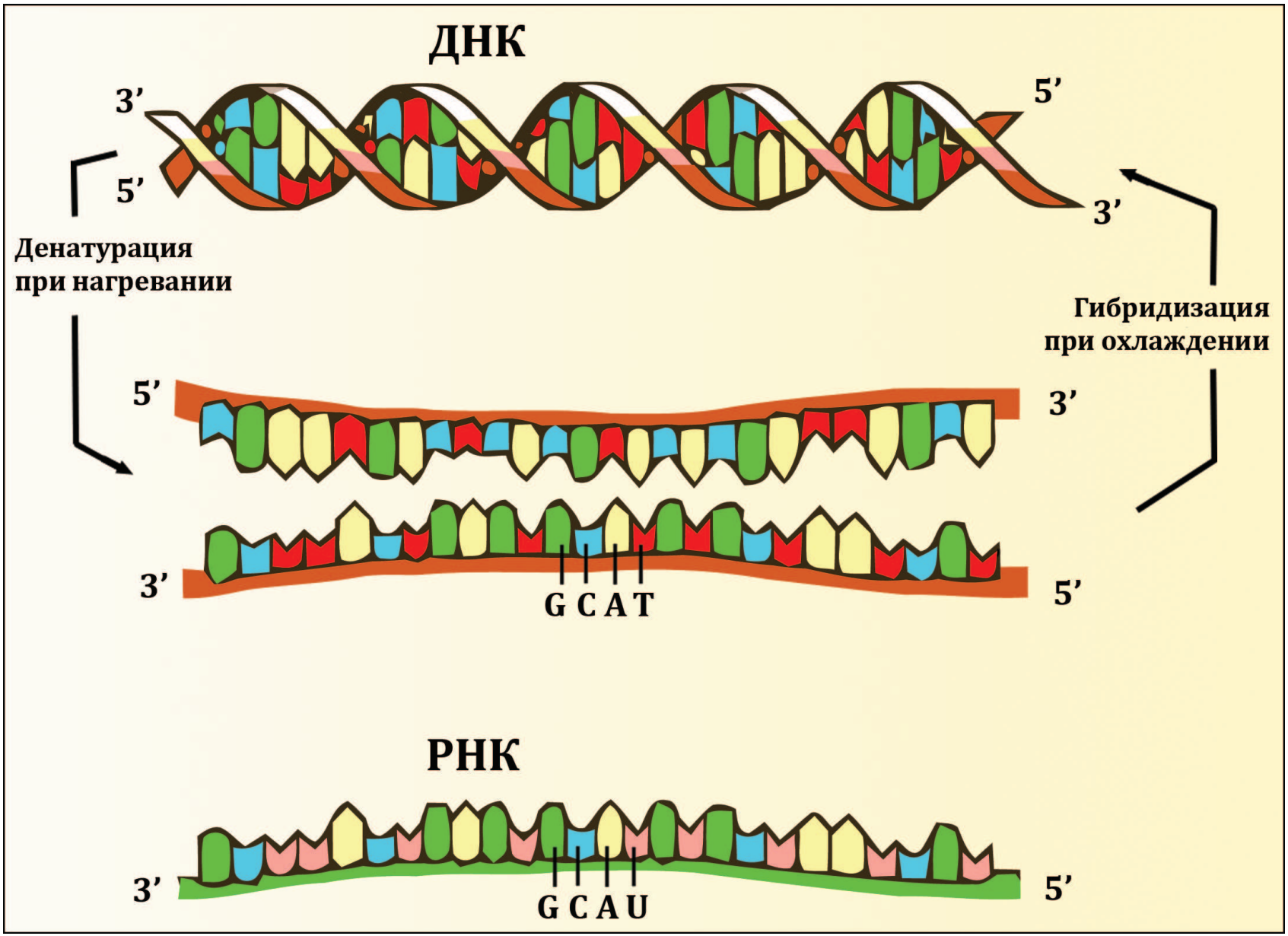


Базовые понятия

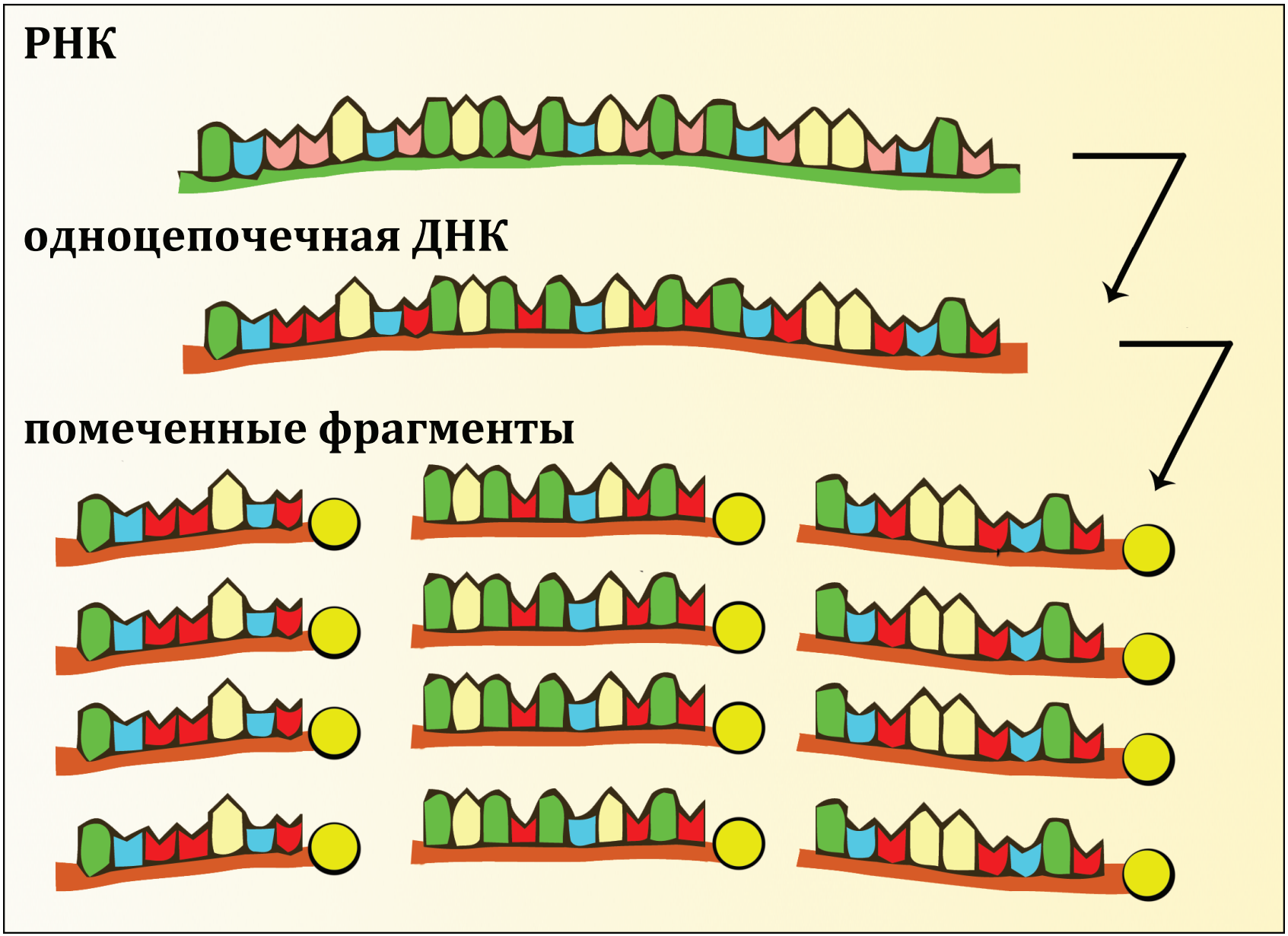
ДНК — молекула, содержащая информацию, необходимую для функционирования клетки.
Ген — участок ДНК, несущий какую-либо целостную функциональную информацию.
Экспрессия гена — преобразование информации, содержащейся в гене, в функциональный продукт.
РНК — молекула-посредник, передающий информацию о гене структурам клетки, отвечающим за синтез белка.

Количество молекул РНК в клетке служит мерой активности гена (оценкой экспрессии).

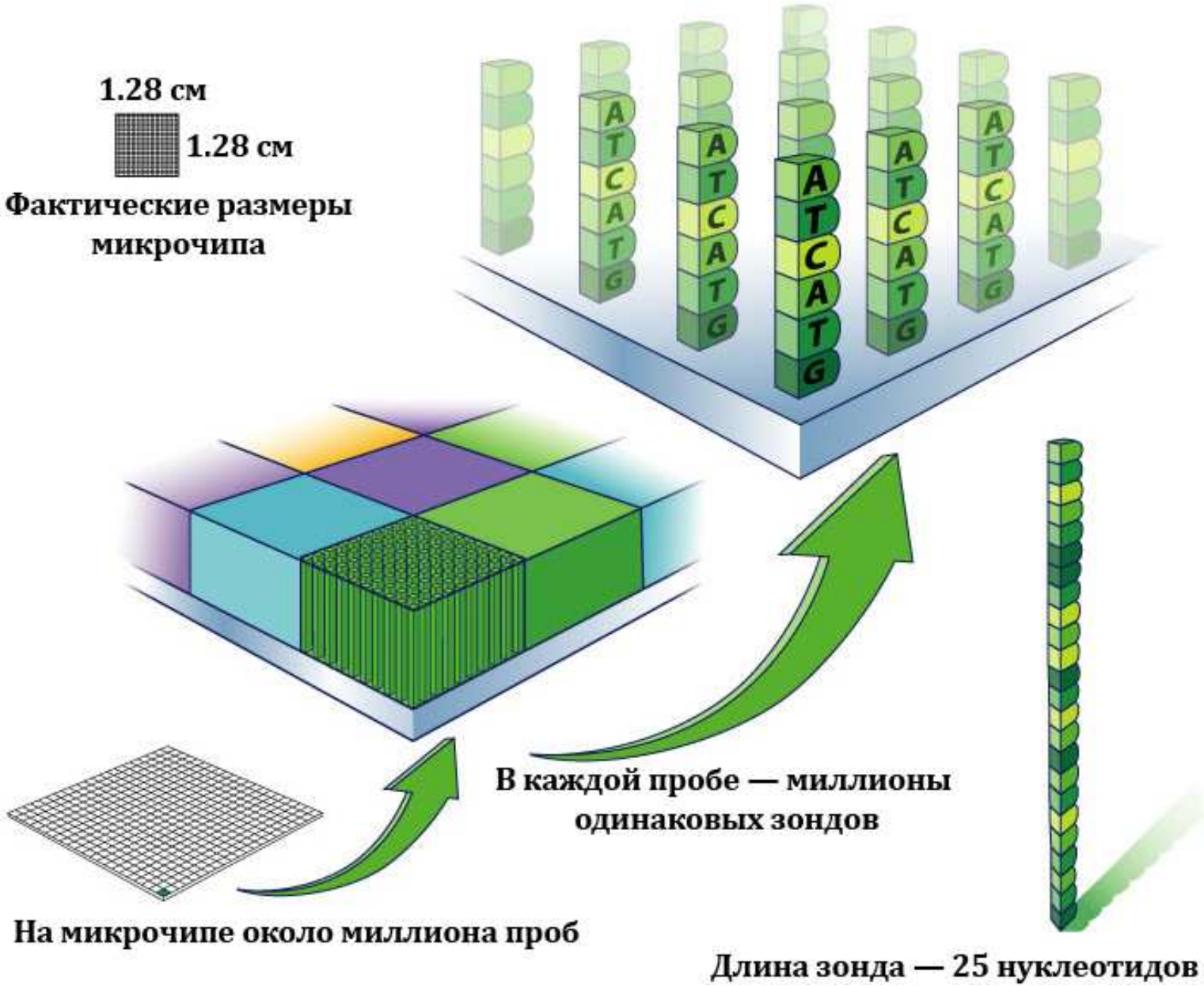
ДНК и РНК


ДНК
3’
5’
5’
3’
Денатурация
при нагревании
Гибридизация
при охлаждении
5’
3’
3’
5’
G C A T
РНК
3’
5’
G C A U

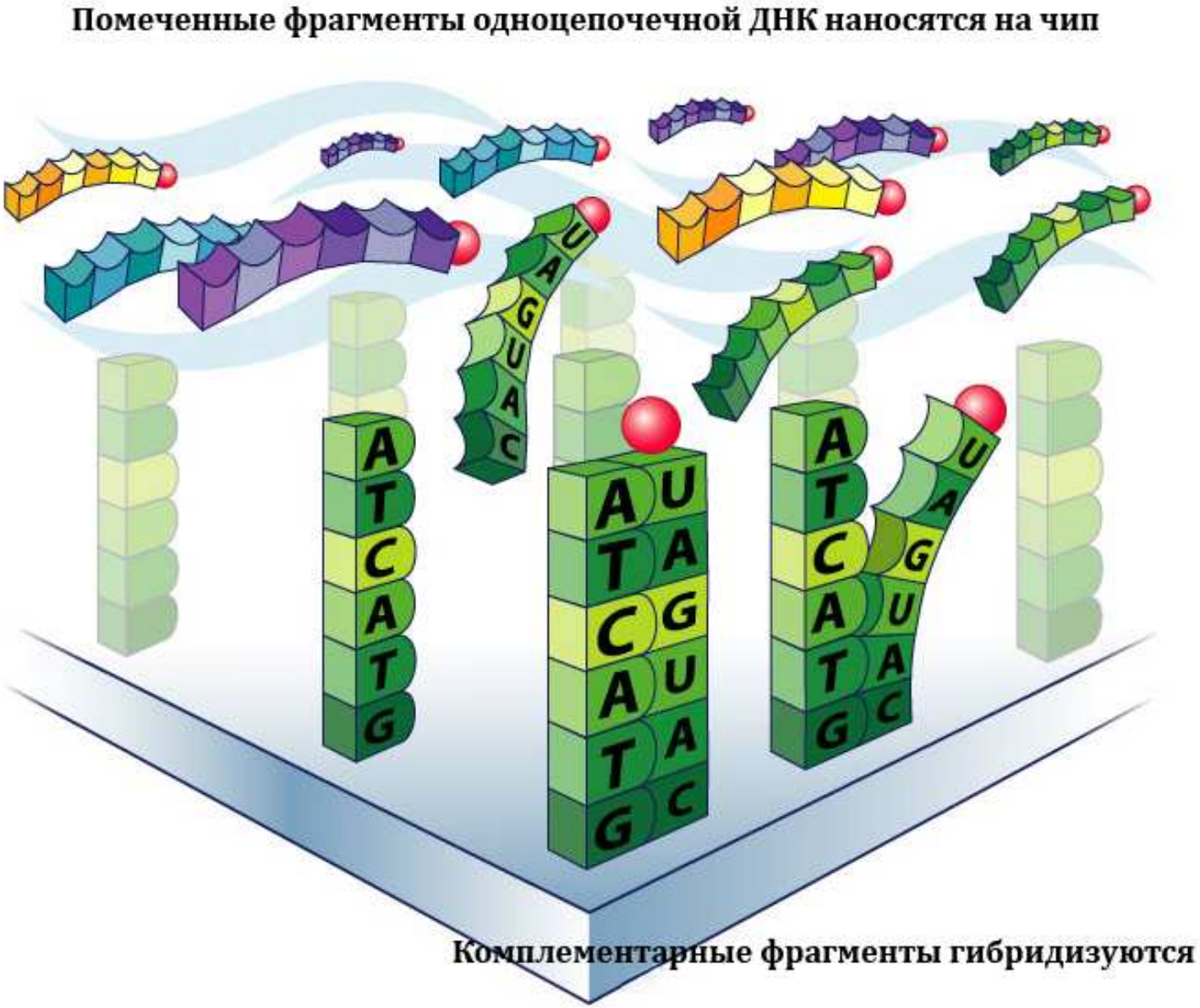
Ход эксперимента



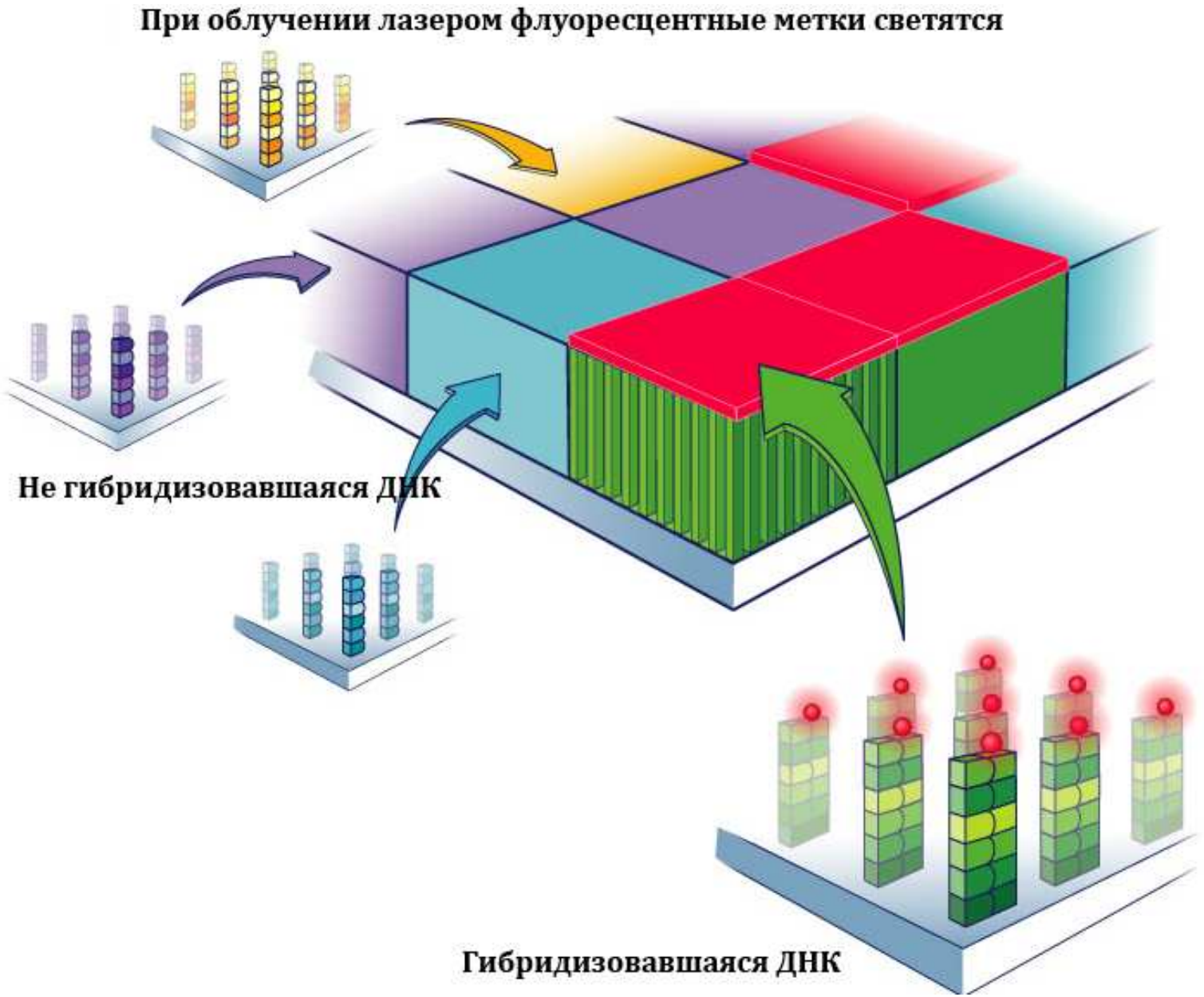
Вид микрочипа



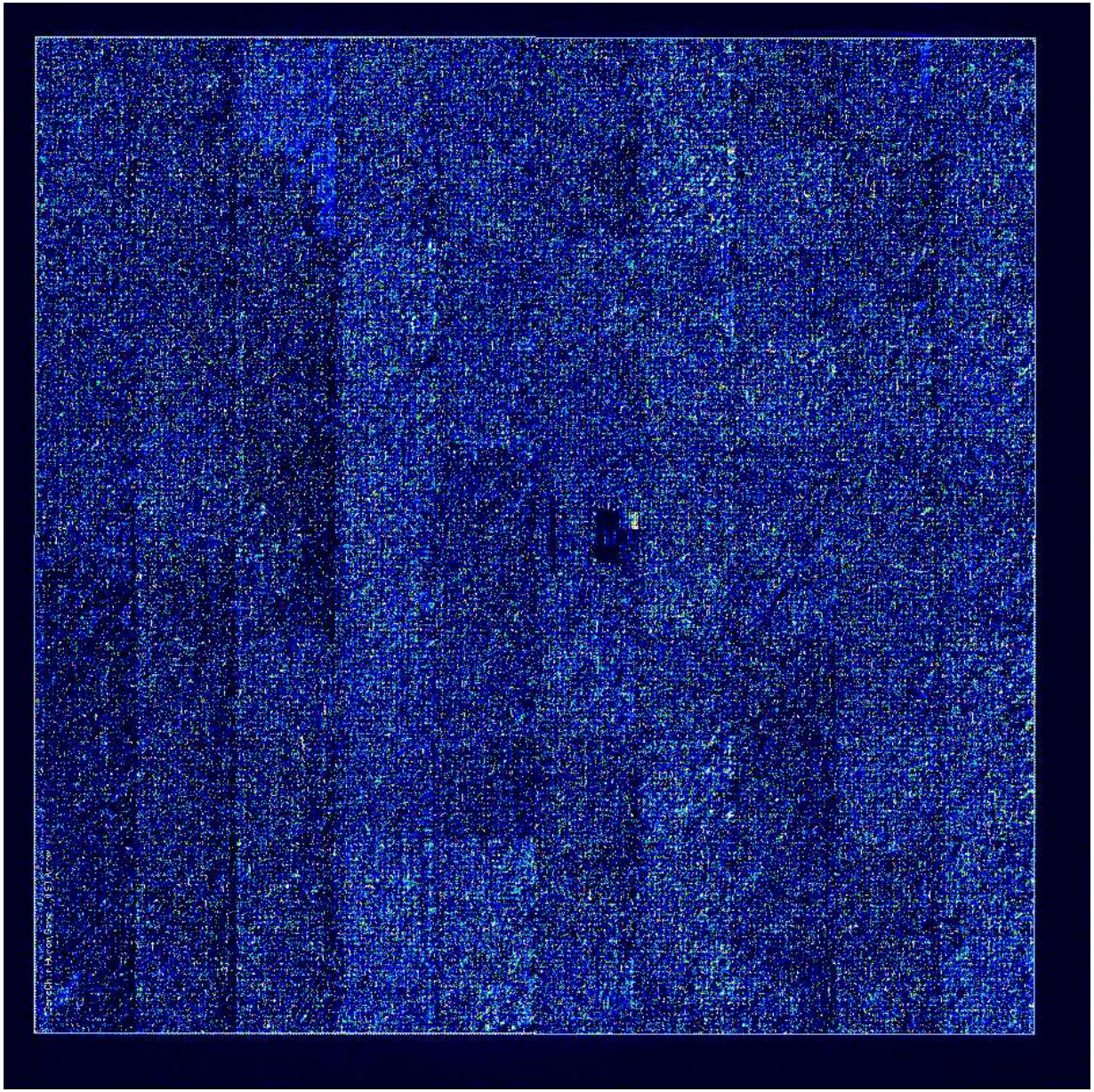
Гибридизация



Сканирование



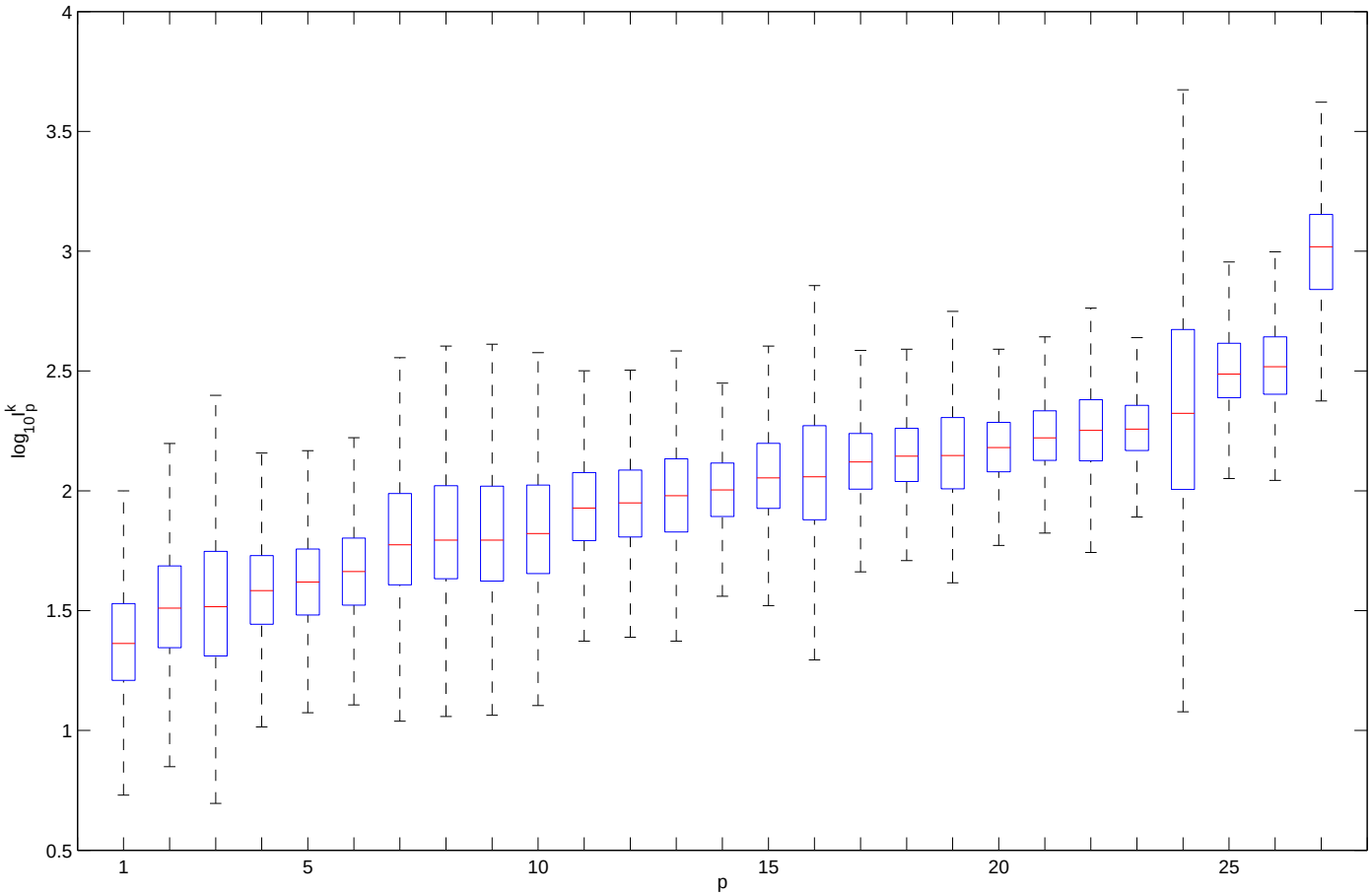
Результат сканирования



Получение оценок экспрессии

1. Изображение со сканера оцифровывается, получаем вектор значений интенсивности флуоресценции проб.
2. Проводится предобработка интенсивностей:
   - фоновая поправка;
   - нормализация.
3. Значения предобработанных интенсивностей всех проб каждого гена **усредняются** (median polish), давая оценку экспрессии.

Распределения интенсивностей проб



Линейная модель, учитывающая степени сродства проб с геном

Известные данные:

I_p^k — интенсивность свечения пробы p на микрочипе k;

$g(p)$ — номер гена, для которого проба p *специфична* (определён конструкцией микрочипа).

Неизвестные параметры:

c_g^k — концентрация РНК гена g на микрочипе k;

a_p — коэффициент *сродства* (affinity) пробы p гену $g(p)$.

$$\hat{I}_p^k = a_p c_{g(p)}^k.$$

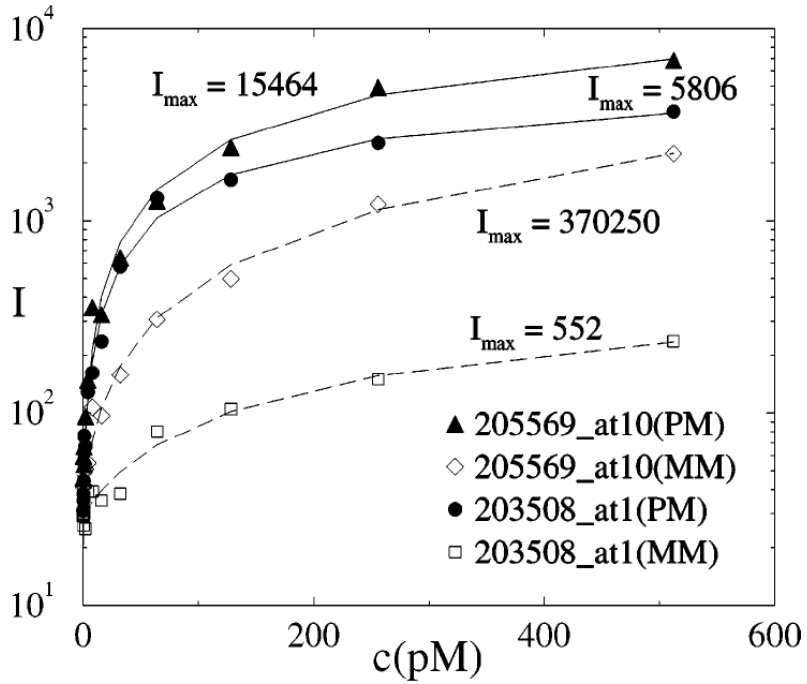
- ограничений $=$ $\underset{735\,497}{\text{проб}}$ $\times$ чипов.
- неизвестных $=$ $\underset{735\,497}{\text{проб}}$ $+$ $\left(\underset{26\,902}{\text{генов}}\right)$ $\times$ чипов.

Необходимо иметь хотя бы два микрочипа.

Недостатки стандартных методов

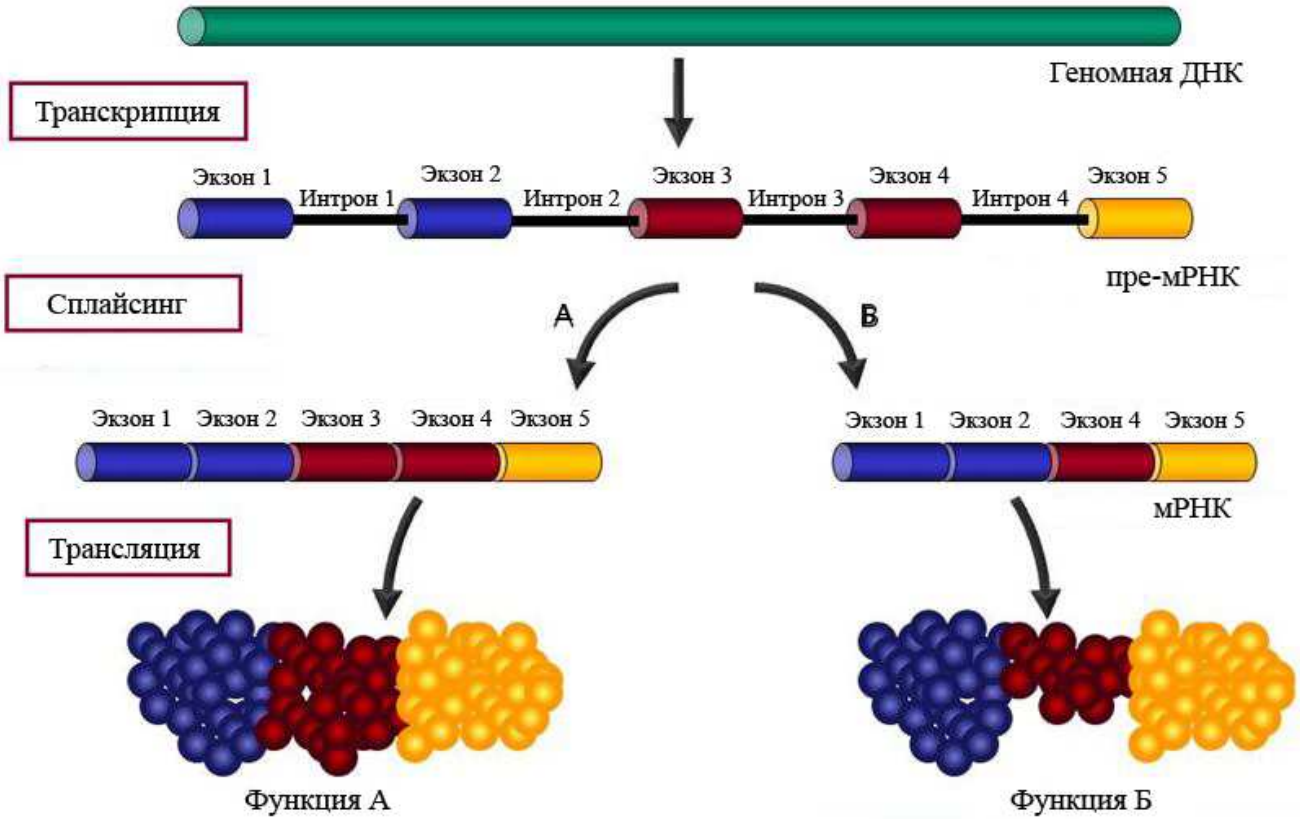
- Коэффициенты сродства не фиксированы, а определяются по анализируемой выборке.
- Зависимость интенсивностей свечения от концентраций РНК гена лучше описывается функцией Лэнгмюра:

$$\hat{I}_p^k = \frac{a_p c_{g(p)}^k}{1 + b_p c_{g(p)}^k}$$

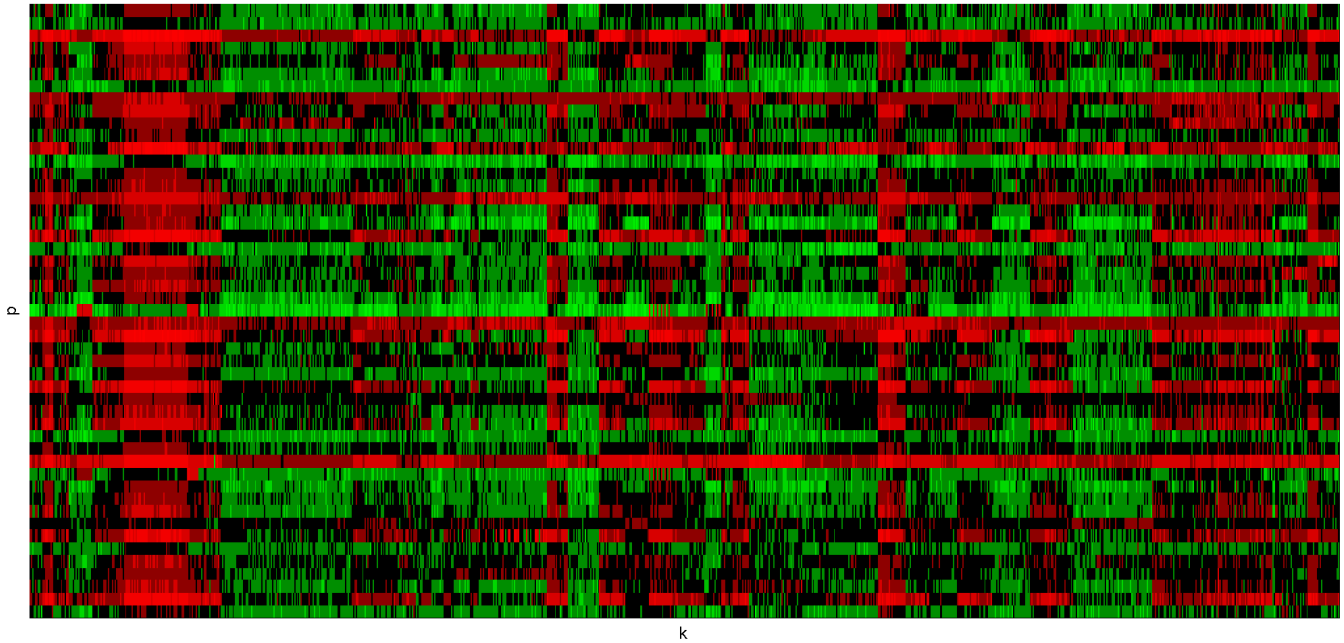


Недостатки стандартных методов

- Не учитывается эффект альтернативного сплайсинга.



Эффект альтернативного сплайсинга



Постановка задачи

Задача: построить метод настройки параметров нелинейной модели Лэнгмюра с учётом альтернативного сплайсинга.

Имеющиеся данные: интенсивности флуоресценции 735497 проб на 3459 микрочипах Affymetrix Human Gene 1.0 ST; данные взяты из общедоступной базы данных NCBI Gene Expression Omnibus.

Параметры модели будем находить как результат минимизации некоторой функции потерь

$$D\left(I, \frac{ac}{1+bc}\right) \to \min_{a,b,c}.$$

Функция потерь

Варианты выбора функции потерь:

- норма Фробениуса (оптимальна для аддитивного гауссовского шума)

$$D(P,Q) = \sum_{i,j} (p_{ij} - q_{ij})^2;$$

- норма l_1 (устойчива к выбросам, оптимальна для аддитивного лапласовского шума)

$$D(P,Q) = \sum_{i,j} |p_{ij} - q_{ij}|;$$

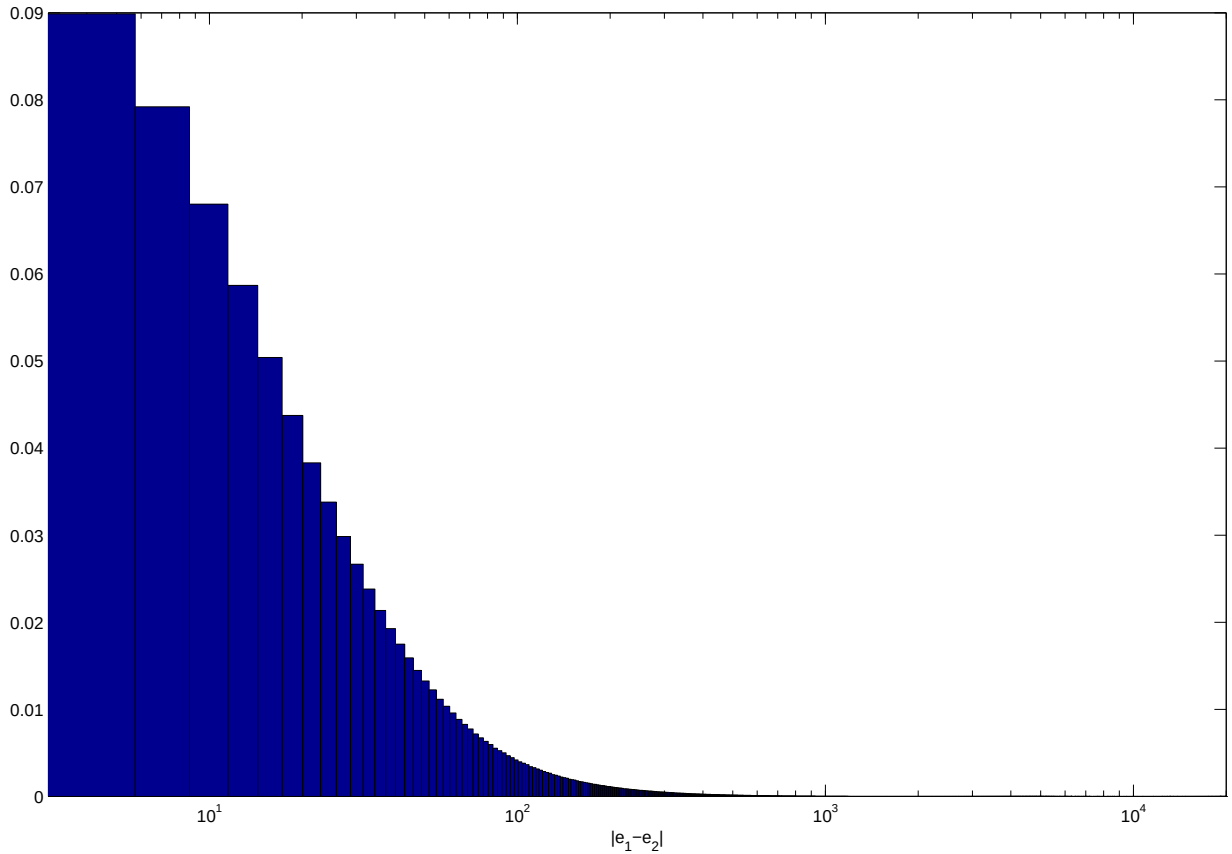
- M-оценки (устойчивы к выбросам)

$$D(P,Q) = \sum_{i,j} \rho(p_{ij} - q_{ij}),$$

$$\rho(x) = \begin{cases} \frac{x^2}{2}, & \text{если } |x| \le 1.345, \\ 1.345\,(|x| - 0.6725), & \text{если } |x| > 1.345; \end{cases}$$

$$\rho(x) = \frac{x^2}{2\,(1+x^2)}.$$

Шум



Плотность распределения модулей разности интенсивностей свечения проб на *технических репликатах* — микрочипах, на которые был нанесён один и тот же образец.

Функция потерь

$$D_{AB}^{(\alpha,\beta)}(P,Q) = \sum_{i,j} d_{AB}^{(\alpha,\beta)}(p_{ij}, q_{ij}),$$

$$d_{AB}^{(\alpha,\beta)}(p,q) = \begin{cases} -\frac{1}{\alpha\beta}\left(p^\alpha q^\beta - \frac{\alpha}{\alpha+\beta}p^{\alpha+\beta} - \frac{\beta}{\alpha+\beta}q^{\alpha+\beta}\right), & \alpha,\beta,\alpha+\beta \neq 0, \\ \frac{1}{\alpha^2}\left(p^\alpha \ln\frac{p^\alpha}{q^\alpha} - p^\alpha + q^\alpha\right), & \alpha \neq 0, \beta = 0, \\ \frac{1}{\alpha^2}\left(\ln\frac{q^\alpha}{p^\alpha} + \left(\frac{q^\alpha}{p^\alpha}\right)^{-1} - 1\right), & \alpha = -\beta \neq 0, \\ \frac{1}{\beta^2}\left(q^\beta \ln\frac{q^\beta}{p^\beta} - q^\beta + p^\beta\right), & \alpha = 0, \beta \neq 0, \\ \frac{1}{2}(\ln p - \ln q)^2, & \alpha = \beta = 0. \end{cases}$$

При различных значениях (α, β) получаемые оценки — ОМП для разных видов распределения шума.

Критерии качества

- Точность приближения:

$$fit\,(I,a,b,c) = \frac{\sum_{k=1}^{K}\sum_{p=1}^{P} I_p^k \left| I_p^k - \frac{a_p c_{g(p)}^k}{1+b_p c_{g(p)}^k} \right| W_p^k}{\sum_{k=1}^{K}\sum_{p=1}^{P} I_p^k W_p^k}$$

W — бинарная матрица, в которой нули соответствуют пробам, не светящимся из-за альтернативного сплайсинга.

Критерии качества

- Воспроизводимость параметров пробы:

$$rep_a = \frac{1}{P}\sum_{p=1}^{P}\left[\frac{|a_{1p} - a_{2p}|}{a_{1p} + a_{2p}} > 0.5\right],$$

$$rep_b = \frac{1}{P}\sum_{p=1}^{P}\left[\frac{|b_{1p} - b_{2p}|}{b_{1p} + b_{2p}} > 0.5\right].$$

Исходная выборка микрочипов разбивается на две части, по каждой из них восстанавливаются параметры проб a_1, b_1 и a_2, b_2. Величины $\frac{|a_{1p}-a_{2p}|}{a_{1p}+a_{2p}}, \frac{|b_{1p}-b_{2p}|}{b_{1p}+b_{2p}}$ принимают значения на $[0, 1]$; их распределение имеет два выраженных пика в нуле и единице, поэтому будем усреднять индикаторы $[x > 0.5]$.

Критерии качества

- Воспроизводимость оценок экспрессии:

$$rep_c = \frac{1}{5GK} \sum_{k=1}^{K} \sum_{p=1}^{P} \sum_{i=1}^{5} \frac{\left| c_g^k - c_{g,\bar{p}_i}^k \right|}{c_g^k + c_{g,\bar{p}_i}^k}$$

Для каждого гена g исключим из рассмотрения пробу p_i и оценим концентрацию РНК гена по множеству оставшихся проб $P(g) \setminus p_i$; $c_{g,\bar{p}_i}$ — полученный вектор оценок экспрессии. Мера воспроизводимости — расстояние между c_g и $c_{g,\bar{p}_i}$; для большей устойчивости усредняется по пяти разным пробам p_i.

Оптимизационная задача

$$\sum_{k=1}^{K}\sum_{p=1}^{P} d_{AB}^{(\alpha,\beta)}\left(I_p^k, \frac{a_p c_{g(p)}^k}{1+b_p c_{g(p)}^k}\right) \to \min_{a,b,c},$$

$$a_p \geq 0, b_p \geq 0, \quad p = 1,\ldots,P,$$

$$c_g^k \geq 0, \quad g = 1,\ldots,G, \quad k = 1,\ldots,K.$$

Благодаря сепарабельности функции потерь задача распадается на G независимых подзадач:

$$\sum_{k=1}^{K}\sum_{p\in P(g)} d_{AB}^{(\alpha,\beta)}\left(I_p^k, \frac{a_p c_g^k}{1+b_p c_g^k}\right) \to \min_{a_g,b_g,c_g}, \tag{1}$$

$$a_p \geq 0, b_p \geq 0, \quad p \in P(g),$$

$$c_g^k \geq 0, \quad k = 1,\ldots,K.$$

Оптимизационная задача

Пусть a_g, b_g, c_g — решение задачи (1), тогда для произвольной константы $C > 0$ векторы $\frac{1}{C} \cdot a_g, \frac{1}{C} \cdot b_g, C \cdot c_g$ тоже будут являться решением.

Для однозначности решения добавим условие нормировки, используемое в стандартных методах обработки микрочиповых данных:

$$\prod_{p \in P(g)} a_p = 1, \quad g = 1, \ldots, G.$$

Для обеспечения устойчивости оценок экспрессии добавим к функции потерь регуляризующее слагаемое

$$\frac{\alpha_c}{2} \sum_{k=1}^{K} \left(c_g^k \right)^2,$$

где α_c — параметр регуляризации.

Оптимизационная задача

$$f\left(I_g, a_g, b_g, c_g, \alpha_c\right) = \sum_{k=1}^{K} \sum_{p \in P(g)} d_{AB}^{(\alpha,\beta)} \left(I_p^k, \frac{a_p c_g^k}{1 + b_p c_g^k} \right) + \frac{\alpha_c}{2} \sum_{k=1}^{K} \left(c_g^k\right)^2 \to \min_{a_g, b_g, c_g},$$

$$\prod_{p \in P(g)} a_p = 1, \tag{2}$$

$$a_p \geq 0, b_p \geq 0, \quad p \in P(g),$$

$$c_g^k \geq 0, \quad k = 1, \ldots, K.$$

Для простоты далее индекс g будем опускать.

Решение оптимизационной задачи

Используем метод блочно-покоординатного спуска, делая шаги стандартного метода Ньютона с проецкией на положительную область изменения параметров поочерёдно по a, b и c; благодаря сепарабельности функции потерь каждая из трёх задач минимизации распадается на независимые подзадачи:

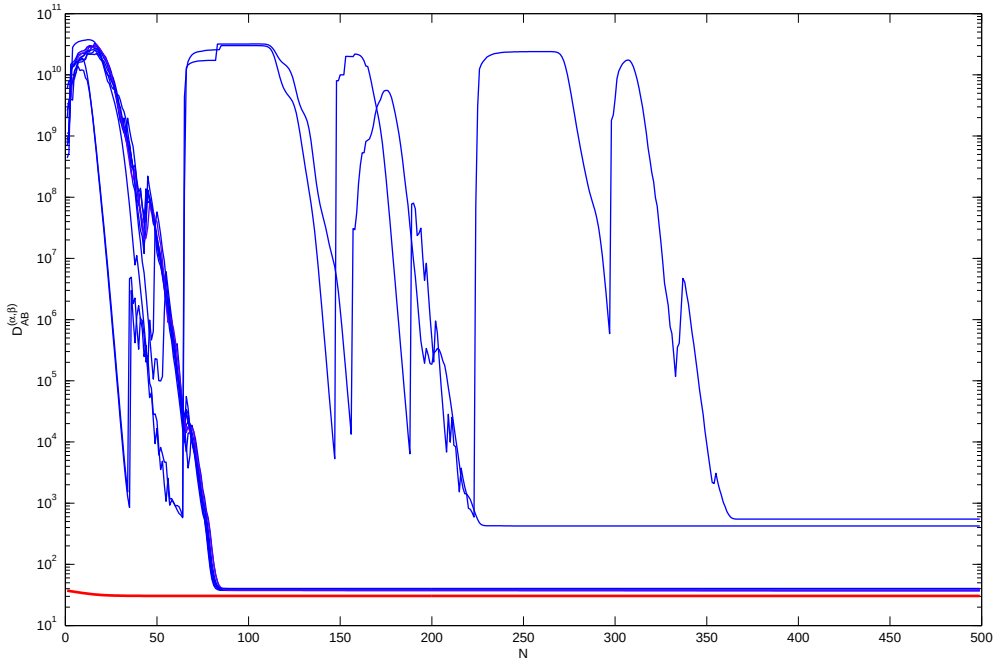
$$a_p = \max\left(0, a_p - \frac{\partial f\left(I, a, b, c, \alpha_c\right)}{\partial a_p} \Big/ \frac{\partial^2 f\left(I, a, b, c, \alpha_c\right)}{\partial a_p^2}\right), \qquad p \in P(g),$$

$$c^k = \max\left(0, c^k - \frac{\partial f\left(I, a, b, c, \alpha_c\right)}{\partial c^k} \Big/ \frac{\partial^2 f\left(I, a, b, c, \alpha_c\right)}{\partial \left(c^k\right)^2}\right), \quad k = 1, \ldots, K,$$

$$b_p = \max\left(0, b_p - \frac{\partial f\left(I, a, b, c, \alpha_c\right)}{\partial b_p} \Big/ \frac{\partial^2 f\left(I, a, b, c, \alpha_c\right)}{\partial b_p^2}\right), \qquad p \in P(g).$$

Начальное приближение

Решение задачи (2) чувствительно к начальному приближению:



В качестве начального приближения для a и c будем брать решение оптимизационной задачи с линейной моделью $\hat{I}_p^k = a_p c^k$, а b инициализируем нулём. Решение находится при помощи мультипликативного алгоритма, предложенного в Cichocki A., Cruces S., Amari S. *Entropy*. 2011. № 13(1). P. 134-170.

Учёт альтернативного сплайсинга

1. $W = \mathbf{1}^{P(g)\times K}$,
2. for iter = 1:maxIter do

$$[a, b, c] = \arg\min_{a,b,c} f\left(I, a, b, c, \alpha_c, W\right),$$

$$E = \frac{1}{I} \cdot \left(\frac{ac}{1+bc}\right) \cdot c;$$

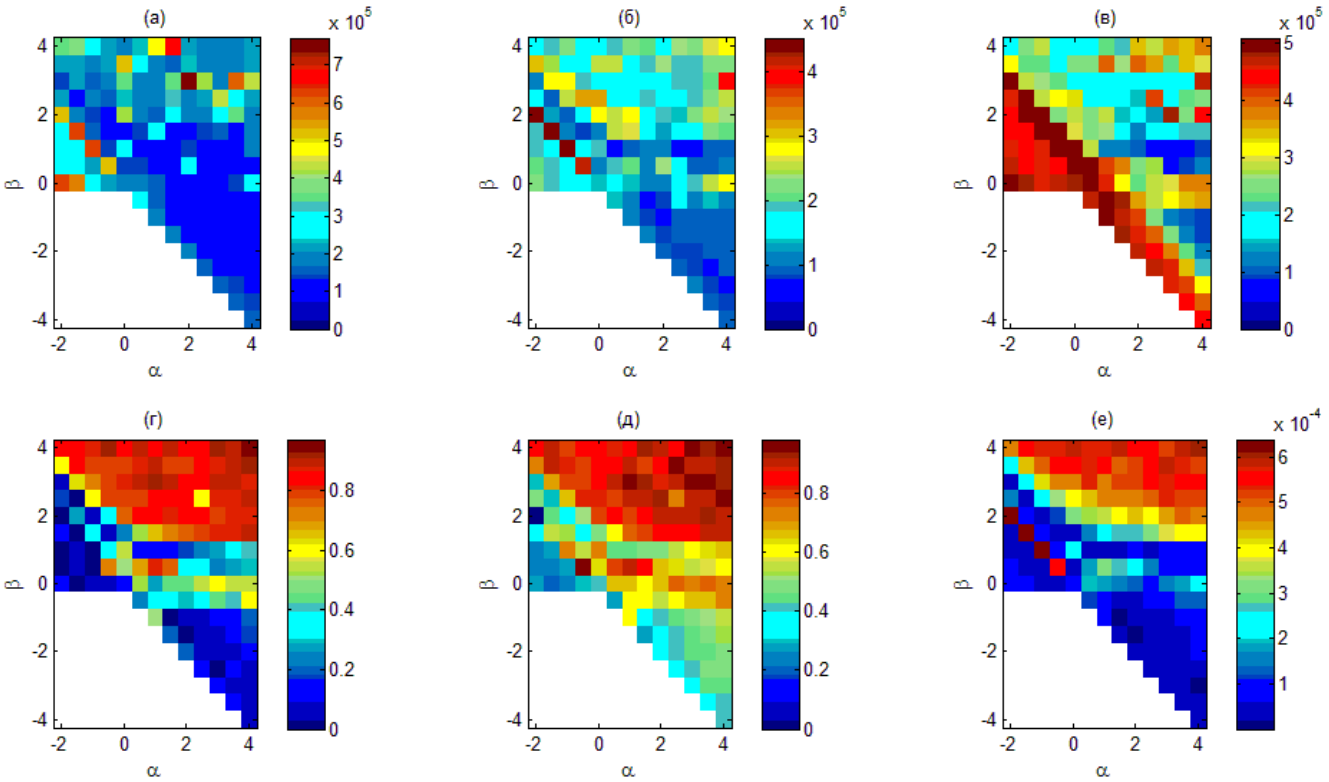
$$W = \left[E <= q_{0.95}(E)\right];$$

3. end for
4. $$err = \frac{\sum_{k=1}^{K}\sum_{p=1}^{P} I_p^k \left| I_p^k - \frac{a_p c_{g(p)}^k}{1+b_p c_{g(p)}^k} \right| W_p^k}{\sum_{k=1}^{K}\sum_{p=1}^{P} I_p^k W_p^k}$$

Схема решения

- на обучающей выборке настраиваем параметры a, b, c;
- зафиксировав a, b, c, на валидационной выборке подбираем оптимальное значение параметра регуляризации α_c;
- на тестовой выборке с фиксированными a, b, c, α_c вычисяем значения критериев качества.

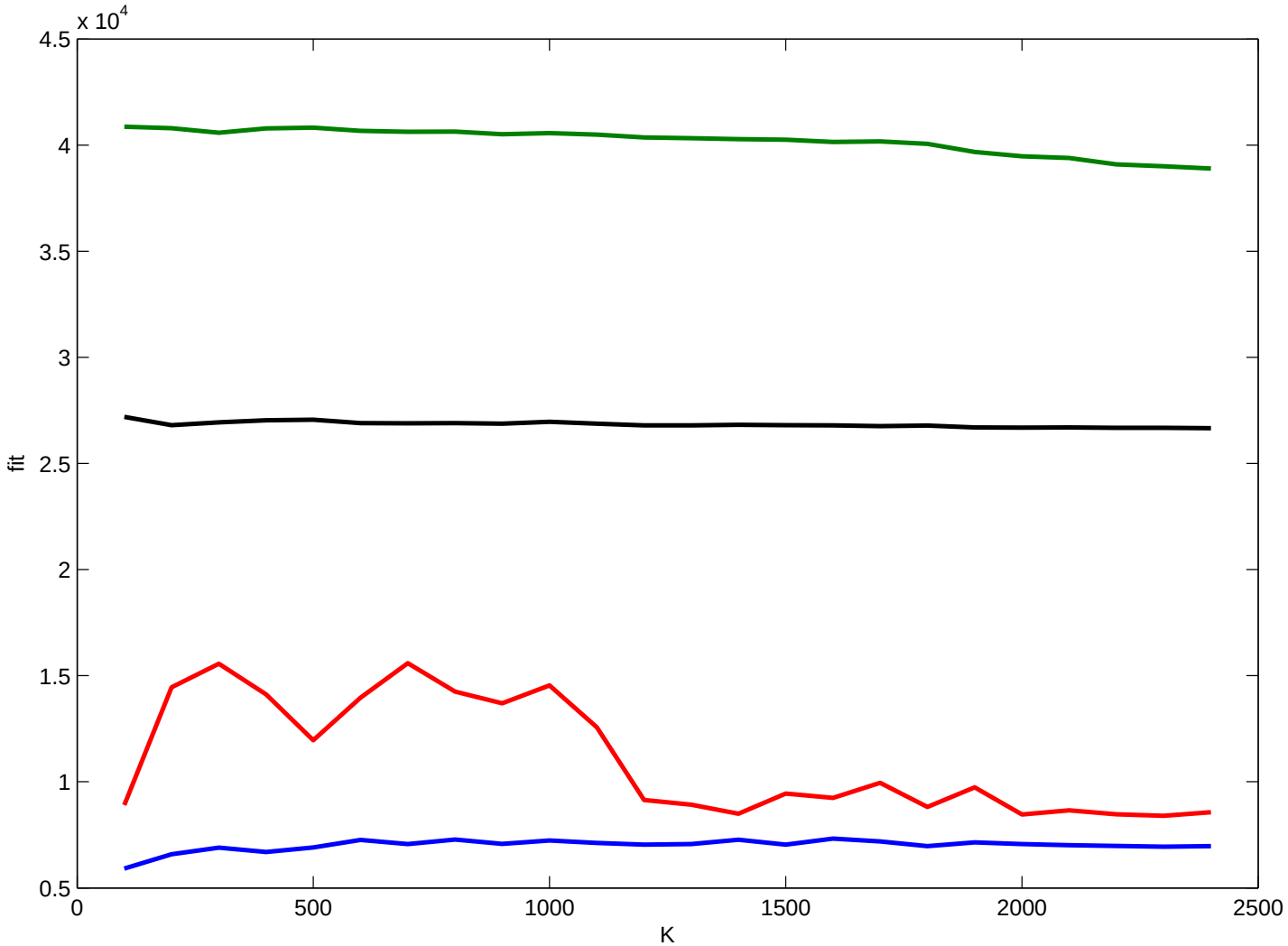
Результат



Верхний ряд: точность приближения, (а) — на обучающей выборке, (б) — на валидационной, (в) — на тестовой; нижний ряд: воспроизводимость, (г) — коэффициентов a, (д) — коэффициентов b, (е) — концентраций.

Результат

Для дальнейших экспериментов выбраны $\alpha = 2, \beta = 1$.
Зависимость точности приближения от размера выборки:



- Разработан оптимизационный метод, позволяющий настроить нелинейную модель интенсивности флуоресценции проб на ДНК-микрочипе.
- На основе настроенной модели создан метод оценивания концентрации РНК генов по микрочиповым данным, позволяющий учесть эффекты насыщения и альтернативного сплайсинга.
- Предложенный метод позволяет точнее объяснить наблюдаемые интенсивности флуоресценции проб, а получаемые с его помощью оценки экспрессии более обладают большей устойчивостью.