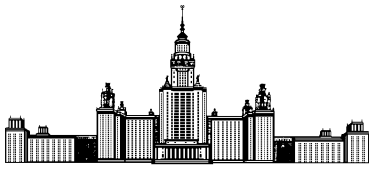
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

ДИПЛОМНАЯ РАБОТА СТУДЕНТКИ 517 ГРУППЫ

«Математическая модель данных микрочипов ДНК и методы оценки её параметров»

Выполнила:

студентка 5 курса 517 группы

Когадеева Мария Сергеевна

Научный руководитель:

д.ф-м.н., доцент

Воронцов Константин Вячеславович

Москва, 2011

Содержание

Аннотация

Анализ данных микрочипов ДНК является одной из перспективных задач современной биоинформатики, однако он осложняется биологическими и техническими вариациями, возникающими на разных стадиях лабораторных экспериментов. В данной работе кратко освещена природа биохимических взаимодействий на поверхности микрочипа и рассмотрены некоторые подходы к их моделированию.

На основании рассмотренных методов предложена модель данных микрочипов ДНК, учитывающая эффекты неспецифических взаимодействий и насыщения. Отличительными особенностями модели являются более подробное рассмотрение механизма неспецифических взаимодействий и учитывание как линейного, так и нелинейного характера зависимости в данных.

Модель построена в рамках подхода к определению абсолютных концентраций молекул в образце, что существенно отличает её от распространённых на сегодняшний день методов, ориентированных на оценку относительных концентраций.

Также исследовано влияние регуляризации на качество настраиваемой модели. Показана устойчивость модели как на этапе настройки параметров, так и при восстановлении концентраций молекул.

Произведено сравнение предложенной модели с существующими методами, показавшее, что оценки относительных концентраций существенно не отличаются. Однако по сравнению с рассмотренными методами модель показала существенно лучшее качество приближения абсолютных концентраций, что делает её применимой для более широкого спектра задач.

Часть I

Введение

1 Биологическая подоплёка

1.1 ДНК и экспрессия генов

Клетка — это минимальная форма жизни. В ядре клетки содержится дезоксирибонуклеиновая кислота (ДНК) — носитель информации, необходимой для функционирования клетки. Молекула ДНК состоит из нуклеотидов четырёх типов: аденин (A), гуанин (G), цитозин (C) и тимин (T). Пространственная структура ДНК была открыта в 1953 г. Уотсоном и Криком. Молекула ДНК сворачивается в двойную спираль по следующему принципу: аденин из одной цепи связывается с тимином другой цепи, а гуанин связывается с цитозином. Этот принцип называется *принципом комплементарности*.

Ген — это участок последовательности ДНК, который контролирует наследуемые свойства организма, благодаря которым организмы отличаются между собой. Генетическая информация, закодированная в ДНК, может быть скопирована в молекулу РНК (рибонуклеиновой кислоты), которая представляет собой одинарную спираль, комплементарную одной из цепочек исходной двойной спирали ДНК. Этот процесс называется *транскрипцией* гена.

В молекуле РНК тимин (T) заменён на урацил (U), который тоже всегда образует пару только с аденином. Есть два основных типа РНК: матричные РНК (мРНК), несущие информацию, и функциональные РНК, которые включают в себя транспортные РНК (тРНК) и рибосомальные РНК (рРНК).

Специальный механизм в клетке считывает информацию с молекулы мРНК и синтезирует определённые молекулы белка. Этот процесс называется *трансляцией* гена.

Гены заключают в себе информацию о том, какой белок, где и в каком количестве будет синтезирован. Центральная догма молекулярной биологии гласит, что ДНК транскрибируется в ядре клетки в молекулу мРНК, которая затем в процессе син-

теза транслируется в белок. Белки и их взаимодействия друг с другом определяют фенотип клетки и организма в целом.

1.2 Измерение экспрессии генов

Чтобы исследовать функцию гена, необходимо знать, какой белок этот ген кодирует, при каких условиях и на каком уровне он активируется. *Экспрессия генов* это процесс, в ходе которого происходит транскрипция и трансляция информации, закодированной в молекуле ДНК (то есть, происходит синтез молекул мРНК и белков). Экспрессию генов можно измерить на двух уровнях: мРНК (что было транскрибировано) и белков (сколько белка синтезировано). Измерить экспрессию генов на уровне белков сложно, поэтому чаще стараются оценить экспрессию на уровне мРНК. Методы оценки экспрессии генов опираются на предположение о том, что если в образце присутствуют молекулы мРНК, соответствующие определённым генам, то эти гены экспрессируются и контролируют синтез белков. Таким образом, ген считается экспрессированным, если кодирующая его ДНК транскрибируется в комплементарную мРНК, и оценкой экспрессии генов является концентрация молекул мРНК в образце.

Существует несколько способов измерить экспрессию генов на уровне концентраций мРНК. Среди них — секвенирование, полимеразная цепная реакция и микрочипы ДНК.

1.2.1 Секвенирование РНК

Секвенирование РНК это одна из самых последних разработок в области молекулярной биологии. Секвенирование РНК (RNA-Seq) относится к методам оценки состава образца, базирующимся на определении нуклеотидных последовательностей молекул. Хотя технология секвенирования все ещё находится в стадии разработки, уже сейчас этот метод обладает рядом существенных преимуществ [28].

Во-первых, RNA-Seq не требует априорных знаний о составе образца и возможных последовательностях генов. Во-вторых, секвенирование обладает очень низким уровнем шума. В-третьих, метод не имеет верхних границ диапазона измерений, таким образом позволяет измерять широкий спектр уровней экспрессии генов. Результаты

секвенирования обладают хорошей воспроизводимостью, что было показано как на технических, так и на биологических репликатах. Marioni et al. [23] провели серию экспериментов, показавших высокую воспроизводимость результатов секвенирования на платформе Illumina.

Однако этот метод обладает также рядом недостатков. На сегодняшний момент использование секвенаторов для определения концентрации РНК в образце осложняется техническими требованиями к подготовке образца и проведению эксперимента. Секвенирование хорошо подходит для определения наличия или отсутствия определенного гена в образце, но количественную оценку концентрации получить несколько сложнее. К тому же, секвенирование может стоить в 10-100 раз дороже, чем другие методы, такие, например, как микрочипы ДНК.

1.2.2 Количественная ПЦР

Количественная ПЦР (полимеразная цепная реакция) позволяет одновременно измерить количество определённой молекулы ДНК в образце и увеличить его в несколько раз. Помимо технических погрешностей основным недостатком ПЦР является то, что одновременно можно определить количество только одного специализированного гена. Поэтому этот метод часто используется для проверки результатов, полученных в ходе определения концентрации с помощью других методов. Для этого отбирают небольшое количество наиболее интересных для исследования генов и сравнивают оценки их экспрессии с помощью микрочипов и количественной ПЦР.

1.2.3 Микрочип ДНК

Технология микрочипов ДНК позволяет получить оценку экспрессии десятков тысяч генов одновременно всего за несколько часов. С помощью микрочипов можно изучать мРНК, полученные из разных тканей при разных состояниях клеток. При этом можно сравнить экспрессию генов в больных и здоровых или обработанных лекарством клетках и определить, какие гены и при каких экспериментальных условиях активируются и могут влиять на процессы, происходящие в клетке. Таким же образом можно использовать микрочипы для определения активных генов в разных тканях на разных стадиях развития эмбриона.

Основной принцип работы микрочипов ДНК заключается в том, что на поверхности микрочипа закреплены на известных позициях определённые участки одинарных цепочек ДНК. Исследуемый образец специально готовят таким образом, чтобы в нём находились одинарные цепочки мРНК. Согласно принципу комплементарности одинарные цепочки в образце вступят в реакцию с участками цепочек, закреплённых на микрочипе. После этого на образец можно нанести флюоресцентные метки и по результатам сканирования микрочипа определить, какие именно участки цепочек ДНК вступили в реакцию, и оценить их концентрацию.

Однако определение экспрессии генов с помощью микрочипов ДНК обладает рядом недостатков и ограничений. Во-первых, на этапе сканирования может возникнуть фоновый шум и засветка, обусловленные погрешностью сканера. Во-вторых, пробы существенно отличаются по характеристикам взаимодействия со свободными цепочками мРНК в образце. Возникает биологический шум, связанный с тем, что на микрочипе могут происходить побочные реакции, влияющие на интенсивность свечения проб.

С помощью микрочипов можно сравнить уровни экспрессии генов на разных стадиях эксперимента, однако по данным из одной выборки надёжно оценить концентрацию РНК определённых генов сложнее. К тому же с помощью микрочипов можно оценивать экспрессию только тех генов, для которых предусмотрены пробы — комплементарные участки ДНК на чипе.

Как бы то ни было, микрочипы ДНК выигрывают у других методов оценки экспрессии генов по совокупности параметров качества измерений, сложности проведения эксперимента и финансовых затрат.

Основной целью данной работы является анализ данных микрочипов ДНК, поэтому устройство и принципы работы последних будут рассмотрены подробнее.

Устройство микрочипа *Микрочип Affymetrix* представляет собой кварцевую пластину, на которую с помощью фотолитографии нанесены *пробы* — олигонуклеотидные последовательности длиной 25 нуклеотидов. Пробы связаны с молекулами поверхности короткими ковалентными связями на расстоянии примерно 10 нм друг от друга. Поверхность микрочипа делится на сотни тысяч областей, называемых *признаками(features)*, размер которых обычно составляет от 11 до 20 мк. В каждой

области расположены пробы с известной последовательностью нуклеотидов, способные связывать определённый участок одинарной цепочки мРНК(ДНК) (Рис. 1).

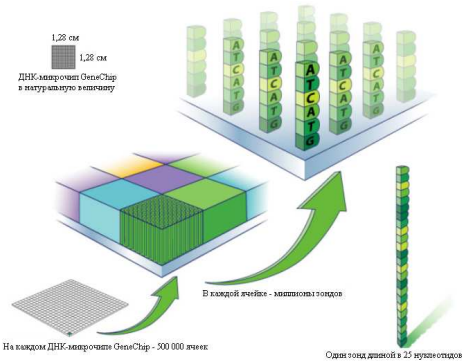


Рис. 1: Схема устройства микрочипа ДНК

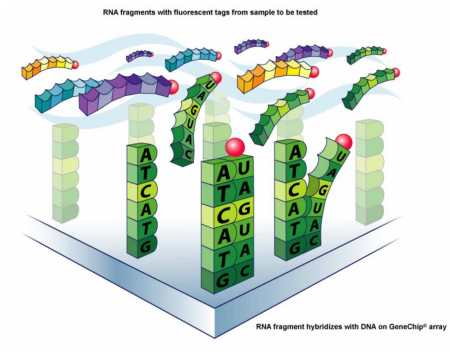


Рис. 2: Гибридизация и окрашивание

Принципы работы микрочипа Ключевым моментом определения экспрессии генов с помощью микрочипов является реакция гибридизации (связывания) проб с последовательностями кодирующих РНК из исследуемого образца длиной от 50 до 100 нуклеотидов. После гибридизации на участки цепочек РНК наносятся флюоресцентные метки, затем микрочип сканируют и определяют интенсивность свечения проб (Рис. 2). Каждому гену соответствует набор из 11-20 специфичных проб (количество зависит от модели микрочипа) (Рис. 3). Для некоторых моделей чипов каждая проба представляет собой пару PM (Perfect Match) и MM (MisMatch) проб, где PM состоит из 25 нуклеотидов, комплементарных соответствующему гену, а MM идентична PM-пробе за исключением серединного (13-го) основания, которое заменено на комплементарное. По интенсивности свечения проб оценивают количество успешных гибридизаций проб с участками РНК, концентрацию РНК в образце и, таким образом, экспрессию соответствующих генов.

Разработка проб Если специфичность гибридизации пробы можно предсказать по её составу, то эффективность связывания короткой (длиной 25 нуклеотидов) пробы с длинным участком гена предсказать гораздо сложнее. Большинство проб, предназначенных для гибридизации с определёнными генами, не будут эффективно связываться с длинной молекулой РНК. Некоторые пробы обладают низкой эффектив-

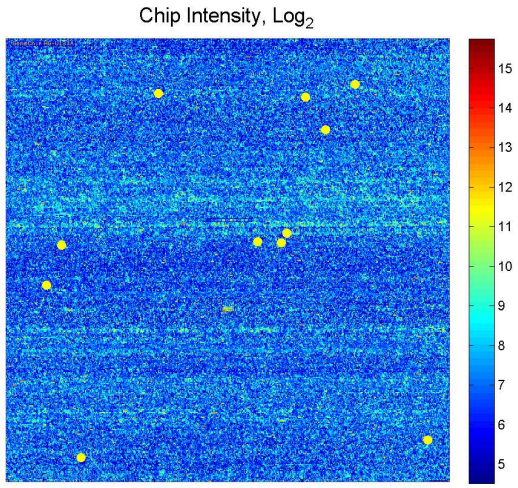


Рис. 3: Результат сканирования микрочипа ДНК. Желтыми точками обозначены пробы, относящиеся к одному гену. Интенсивность свечения представлена в логарифмической шкале.

ностью связывания и, как следствие, интенсивностью свечения, что мешает оценивать гибридизацию и концентрацию генов в образце.

Отбор проб, обладающих высокой интенсивностью связывания с участками специфичных генов, является одной из главных задач при конструировании микрочипа. Чтобы предсказать интенсивность гибридизации пробы, разработчики Affymetrix используют компьютерные модели, базирующиеся на результатах нескольких тысяч экспериментов, в которых наблюдались реакции взаимодействия в сложных смесях. Успех этого подхода, несомненно, отражается на популярности микрочипов Affymerix. Однако существуют и другие подходы к выявлению наиболее эффективных проб. Luebke et al. [22] сделали несколько чипов с пробами к небольшому набору генов и определили наиболее эффективные пробы, измеряя интенсивность свечения. Однако этот метод не позволяет учесть эффекты неспецифических взаимодействий, к тому же он неприменим для большого числа генов. О других методах отбора эффективных проб будет сказано немного позднее.

2 Анализ данных микрочипов ДНК

Основной задачей анализа данных микрочипов ДНК является определение уровня экспрессии генов в образце по интенсивности свечения проб на чипе. При анализе

данных возникает необходимость учитывать большое количество факторов и параметров, влияющих на интенсивность свечения проб. Корреляция между интенсивностями двух проб одного ДНК-микрочипа, представляющих один и тот же ген, обычно превышает 95%. Часто этот факт интерпретируют как подтверждение хорошей воспроизводимости экспериментов с чипами. Однако если один и тот же биологический образец нанести на разные микрочипы, корреляция между полученными интенсивностями может лежать в диапазоне от 60 до 80%. Корреляция интенсивностей свечения проб на чипах с образцами, полученными от разных особей при одинаковых условиях, может быть ещё ниже. Если эксперименты проводятся в разных лабораториях, корреляция между их результатами может быть совсем низкой.

Такая низкая воспроизводимость интенсивностей вызвана большим количеством биологических и технических факторов. Основными типами вариации являются

- *Биологическая вариация*, которая обусловлена особенностями протекания химических реакций на поверхности микрочипа;
- *Техническая вариация*, возникающая на этапе выделения образцов, их окрашивания и гибридизации. Нейтрализация эффектов технической вариации и ошибки измерения производится на этапе предобработки данных микрочипов ДНК;
- *Погрешность измерения*, вызванная особенностями сканера, например, светочувствительностью или наличием пыли внутри сканера.

2.1 Техническая вариация

Техническая вариация возникает на этапах подготовки и проведения эксперимента и сканирования микрочипа. Данные микрочипов ДНК, полученные для одинакового эксперимента, проведенного при разных условиях, в разных лабораториях, разными исследователями, могут очень сильно отличаться. Кроме того, при изготовлении микрочипов невозможно добиться абсолютной идентичности, микрочипы могут немного отличаться друг от друга по характеристикам гибридизации проб или чувствительности к флюоресцентным меткам. Сканер тоже может внести погрешности при обработке данных: разные микрочипы могут иметь разный спектр

интенсивностей свечения проб. Иногда данные отличаются вследствие неравномерного распределения образца по поверхности чипа или попадания в образец пузырька воздуха.

Одной из основных задач анализа данных микрочипов ДНК является сравнение уровней экспрессии генов на разных стадиях эксперимента. Поэтому большое значение имеет *предобработка данных*, которая состоит из трёх этапов: *фоновой поправки, нормализации и суммаризации*. Фоновая поправка направлена на устранение шумовых эффектов и особенностей сканирования, нормализация помогает привести данные нескольких микрочипов к виду, пригодному для сравнения, а на этапе суммаризации из данных об интенсивности свечения проб получают оценки экспрессии генов. Рассмотрим подробнее каждый из этих этапов.

2.1.1 Фоновая поправка

Фоновая поправка — важный этап предобработки данных микрочипов ДНК. Необходимость фоновой поправки обуславливается такими источниками технической вариации, как погрешность измерений оптического прибора, светимость подложки сканера, наличие пузырьков воздуха в образце и т.д.

Для анализа фонового шума изначально предполагалось использовать MM-пробы, отличающиеся от PM-проб серединным нуклеотидом. По задумке разработчиков микрочипа MM-пробы не должны были вступать в реакцию со специфическими генами, поэтому с их помощью можно было бы оценить фоновую интенсивность. Этот подход сразу же продемонстрировал свою несостоятельность: оказалось, что в среднем для микрочипа интенсивность около 30% MM-проб превышает интенсивность соответствующих им PM-проб. Таким образом, вычитание интенсивностей MM-проб приводит к бессмысленному результату, поскольку экспрессия гена оказывается отрицательной.

Существуют более гибкие алгоритмы учёта фоновой поправки, не использующие интенсивность MM-проб или использующие её лишь частично. Ниже описаны наиболее популярные из них.

- **Ideal Mismatch**

  Чтобы нейтрализовать эффект отрицательных значений экспрессии при вычи-

тании интенсивности MM-проб, компанией-изготовителем микрочипов Affymetrix была разработана концепция Ideal Mismatch. Идея заключается в том, что в случае отрицательности разницы интенсивностей PM и MM-проб из интенсивности PM-пробы вычитается некоторая величина, меньшая интенсивности MM-пробы.

Для каждого набора проб, соответствующих одному гену, вычисляется значение фоновой интенсивности SB, представляющее собой одношаговое взвешенное среднее Тьюки по множеству логарифмов отношений PM-интенсивностей к MM-интенсивностям в каждой паре проб. Если i — номер пробы, а k — номер набора, то фоновая поправка I_0 вычисляется следующим образом:

$$I_0 = \begin{cases} MM_i^{(k)}, & MM_i^{(k)} < PM_i^{(k)}, \\ \dfrac{PM_i^{(k)}}{2^{SB_k}}, & MM_i^{(k)} \geqslant PM_i^{(k)},\ SB_k > \tau_c, \\ \dfrac{PM_i^{(k)}}{2^{\tau_c/(1+(\tau_c - SB_k)/\tau_s)}}, & MM_i^{(k)} \geqslant PM_i^{(k)},\ SB_k \leqslant \tau_c, \end{cases}$$

$$SB_k = T_{bk}\left(\log_2 \frac{PM_i^{(k)}}{MM_i^{(k)}},\ i = 1, \ldots, n_k\right).$$

Здесь $MM_i^{(k)}$, $PM_i^{(k)}$ — интенсивности MM и PM проб соответственно, T_{bk} — одношаговое взвешенное среднее Тьюки с параметрами $c = 5$, $\epsilon = 0.0001$. Параметр τ_c —константа различия, значение по умолчанию 0.03, τ_s — константа масштабирования, значение по умолчанию 10. Итоговое значение интенсивности для PM-проб с учётом фоновой поправки получается вычитанием из исходных значений интенсивностей PM-проб соответствующей им величины I_0.

- **MAS 5.0 (Affymetrix MicroArray Suite 5.0)**

  Метод MAS 5.0 предложен разработчиками микрочипов Affymetrix для удаления фонового шума. Схема метода следующая: сначала микрочип делят на несколько прямоугольных областей одинаковой площади (по умолчанию 16). Для k-й области фоновая составляющая b_k вычисляется с помощью 2%-квантиля (наименьших значений) интенсивности. Оценивается также дисперсия n_k 2% наименьших интенсивностей. Значение фоновой поправки $b(x, y)$ и её дисперсия $n(x, y)$ для каждой пробы на чипе с координатами (x, y) рассчи-

тываются по формулам

$$b(x,y) = \frac{1}{\sum_{k=1}^{K} \omega_k(x,y)} \sum_{k=1}^{K} \omega_k(x,y) b_k.$$

$$n(x,y) = \frac{1}{\sum_{k=1}^{K} \omega_k(x,y)} \sum_{k=1}^{K} \omega_k(x,y) n_k.$$

Веса $\omega_k(x,y)$ зависят от расстояния между пробой и центрами выделенных областей (рис. 4):

$$\omega_k(x,y) = \frac{1}{d_k^2(x,y) + s_0},$$

где $d_k(x,y)$ — евклидово расстояние между пробой и центром k-го прямоугольника, s_0 — сглаживающий коэффициент (значение по умочанию 100).

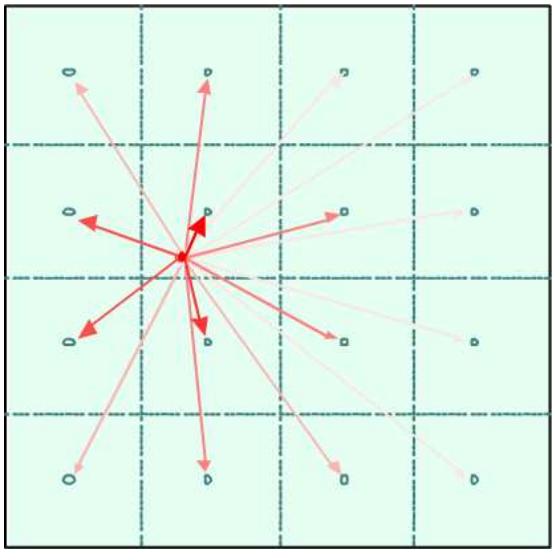


Рис. 4: Схема разбиения микрочипа на области алгоритмом MAS 5.0

Значение интенсивности пробы с координатами (x,y) вычисляется по формуле

$$\hat{I}(x,y) = \max(\max(I(x,y),\, 0.5) - b(x,y),\, NoiseFrac * n(x,y)),$$

где $I(x,y)$ — исходное значение интенсивности, $NoiseFrac$ — доля учитываемой вариации фоновой интенсивности (по умолчанию 0.5).

- **RMA (Robust Multichip Average)**

  В методе учёта фоновой поправки, являющемся частью комплекса RMA методов предобработки данных микрочипов ДНК, используются только интенсивности PM-проб [18]. Интенсивность пробы представляется в виде $PM_{it}^{(k)} =$

$bg_{it}^{(k)} + s_{it}^{(k)}$. Здесь $bg_{it}^{(k)}$ представляет собой фоновый сигнал на микрочипе t, включающий в себя оптический шум и неспецифические взаимодействия (см. Биологическая вариация). Необходимо определить истинный сигнал $s_{it}^{(k)}$. В методе предполагается, что $\mathsf{E}(bg_{it}^{(k)}) = \beta_t$, величина $bg_{it}^{(k)}$ распределена нормально с матожиданием μ и дисперсией σ^2, а $s_{it}^{(k)}$ — экспоненциально со средним α. Оценка сигнала представляет собой условное матожидание $B(PM_{it}^{(k)}) \equiv \mathsf{E}(s_{it}^{(k)}|PM_{it}^{(k)})$.

В обозначениях $a = s - \mu - \sigma^2\alpha$ и $b = \sigma$ оценка сигнала имеет вид:

$$\mathsf{E}(s|PM) = a + b\frac{\phi(\frac{a}{b}) - \phi(\frac{PM-a}{b})}{\Phi(\frac{a}{b}) + \Phi(\frac{PM-a}{b}) - 1},$$

где ϕ и Φ — плотность и функция распределения стандартного нормального распределения соответственно.

- **DFCM (Distribution Free Convolution Model)**

  В рамках данной модели интенсивность проб также представлена в виде $PM_{it}^{(k)} = bg_{it}^{(k)} + s_{it}^{(k)}$. Однако в отличие от алгоритма RMA никаких предположений о распределениях компонент не делается. Алгоритм учёта фоновой поправки состоит из нескольких шагов [9]:

  1. Выделяются наименьшие q_1 процентов значений PM-интенсивностей (обычно доля q_1 достаточно мала и не превышает 30%).
  2. Выделяются наименьшие q_2 процентов (обычно 90% или 95%) значений MM-интенсивностей проб, соответствующих PM-пробам, отобранным на предыдущем шаге. Отобранные значения интенсивностей MM-проб далее служат мерой фонового шума.
  3. С использованием непараметрической оценки плотности распределения шума (как правило, ядерной оценки Епанечникова), ищется мода распределения шума $\hat{\mu}$.
  4. Оценкой стандартного отклонения шума служит $\hat{\sigma}$ — выборочное стандартное отклонения шума со значениями интенсивностей, меньших $\hat{\mu}$, умноженное на $\sqrt{2}$.

5. Значение интенсивности i-й пробы в k-м наборе проб, соответствующих одному гену, рассчитывается по следующей формуле:

$$s_i^{(k)} = \begin{cases} PM_i^{(k)} - \hat{\mu}, & PM_i^{(k)} \geqslant \hat{\mu} + 3\hat{\sigma}, \\ 1 + \big(PM_i^{(k)} - \min\big)\big(\dfrac{2\hat{\sigma} - 1}{\hat{\mu} + 3\hat{\sigma} - \min}\big), & PM_i^{(k)} < \hat{\mu} + 3\hat{\sigma}, \end{cases}$$

где min — минимальное значение интенсивности (PM или MM пробы).

Chen et al. [9] предполагают, что причиной большего значения интенсивности MM-пробы может быть неспецифическая гибридизация. Чтобы контролировать долю неспецифической гибридизации в данных, предлагается настраивать параметр q_1 таким образом, чтобы доля MM проб с интенсивностями, большими, чем у PM проб, составляла 50% для наименьших q_1%. Параметр q_2 выражает вероятность того, что MM проба вступает в неспецифические взаимодействия. Этот параметр в меньшей степени влияет на результат.

DFCM работает лучше других методов фоновой поправки в том случае, если не выполняется гипотеза нормальности распределения шума.

2.1.2 Нормализация

При сравнении нескольких микрочипов из одного или нескольких экспериментов возникают два типа различий: биологические и технические. Биологические различия, то есть изменение концентрации генов в образце, представляют интерес с научной точки зрения. Технические различия возникают на этапе подготовки эксперимента, прокрашивания образцов, вследствие разницы в светочувствительности сканера и т.д. Для выявления биологических различий необходимо исключить или свести к минимуму технические различия, и для этой цели проводят *нормализацию* данных микрочипов. Методы нормализации делятся на две основные группы: методы, использующие чип-эталон, и методы, использующие всю выборку микрочипов. Методы, использующие эталон, делятся в свою очередь на линейные и нелинейные. Среди методов, осуществляющих нормализацию по всей выборке микрочипов без выбора эталона, циклическая локальная регрессия, метод контрастов и квантильная нормализация.

- **Линейная нормализация**

  В стандартном подходе, предлагаемом разработчиками Affymetrix, нормализация применяется к оценкам экспрессии генов. Она осуществляется таким образом, чтобы среднее значение экспрессии на всех чипах было одинаковым. Для этого выбирается эталон — микрочип с медианным значением медианы экспрессии генов. После этого значения экспрессии x_t на каждом чипе t изменяются по формуле

  $$x'_t = \frac{\tilde{x}_{base}}{\tilde{x}_t} x_t,$$

  где $\tilde{x}_t$ — усеченное среднее значений экспресии (наибольшие и наименьшие 2% значений были исключены), x_{base} — значения экспрессий для микрочипа-эталона, x'_t — модифицированное значение экспрессии. Этот подход плохо работает в том случае, когда в данных есть нелинейные зависимости.

- **Нелинейная нормализация**

  Существуют нелинейные методы нормализации. Идея заключается в том, чтобы выбрать подмножество инвариантных проб на чипе-эталоне и остальных чипах и настроить на нём нелинейные функции нормализации $\hat{f}_t(x)$ для каждого чипа t. В качестве функций нормализации могут выступать сглаживающие сплайны, кусочно-линейная медиана, локальная линейная регрессия и т.д. В этом случае значение экспрессий на микрочипе t будет иметь вид

  $$x'_t = \hat{f}_t(x_t).$$

  Отметим, что описанные выше методы применяют как к оценкам экспрессии генов, полученных после суммаризации, так и к интенсивностям свечения проб. Один из главных недостатков этих методов — необходимость выбора эталонного микрочипа.

- **Cyclic Loess (Locally weighted scatterplot smoothing)**

  Метод циклической локальной регрессии (cyclic loess) основан на восстановлении локальной регрессии для величин $M_i = \log_2(\frac{x_{it}}{x_{is}})$ и $A_i = \frac{1}{2}\log_2(x_{it}x_{is})$. Здесь x_{it}, x_{is} — интенсивности свечения i-й пробы на t-м и s-м чипах соответ-

ственно. Пусть $\hat{M}_i$ — значение функции регрессии, тогда поправка нормализации будет равна $M'_i = M_i - \hat{M}_i$, а нормализованные интенсивности проб —

$$x'_{it} = 2^{A_i + \frac{M'_i}{2}}, \qquad x'_{is} = 2^{A_i - \frac{M'_i}{2}}.$$

Для нормализации данных более двух микрочипов выполняется попарная нормализация. В связи с этим недостатком этого метода может являться длительное время работы.

- **Contrast based method**

  Метод контрастов (contrast based method) тоже основан на восстановлении локальной регрессии между величинами M и A. В этом случае восстанавливаются (n-1) кривые нормализации для преобразованных интенсивностей, где n — число микрочипов. Метод контрастов работает быстрее, чем циклическая локальная регрессия, однако тоже может требовать определённых временных затрат.

- **Квантильная нормализация**

  Основной целью метода квантильной нормализации (quantile normalization) является приведение распределений интенсивностей проб на микрочипах к одному распределению. Если n векторов имеют одинаковое распределение, то график квантилей в n-мерном пространстве будет представлять собой прямую, лежащую вдоль диагонали, заданной вектором $(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$. Из этого соображения вытекает, что чтобы привести n векторов к одному распределению, необходимо спроектировать точки n-мерного графика квантилей на диагональ.

  Пусть $q_i = (q_{i1}, \ldots, q_{in})$ — вектор i-квантилей для всех n микрочипов, $d = (\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$ — вектор, задающий диагональ в n-мерном пространстве. Проекция i-квантили на диагональ имеет вид $prod_d q_k = \big(\frac{1}{n}\sum_{j=1}^{n} q_{ij}, \ldots, \frac{1}{n}\sum_{j=1}^{n} q_{ij}\big)$. Это означает, что все интенсивности можно привести к одному распределению, если взять вместо исходных значений среднее i-квантили. Нормализация состоит из следующих этапов:

  1. представить интенсивности n микрочипов в виде матрицы X размера $p \times n$, где p — число проб на чипе;
  2. отсортировать матрицу X по столбцам;

3. посчитать среднее по строкам матрицы X и заменить значение в строке на среднее;

4. отсортировать столбцы матрицы X в первоначальном порядке.

Квантильная нормализация это частный случай преобразования $x'_i = F^{-1}(G(x_i))$, где G оценена с помощью эмпирического распределения интенсивностей каждого мирочипа и F — с помощью эмпирического распределения средних выборочных квантилей. Расширения этого метода могут использовать более гладкие оценки F^{-1} и G.

Один из недостатков этого метода заключается в том, что все квантили обязаны быть равны. Проблема может возникнуть в том случае, когда пробы имеют одинаковые значения интенсивностей на всех чипах. Однако на практике концентрация гена определяется сразу по нескольких пробам из подмножества, соответствующего этому гену, поэтому эта проблема оказывается незначительной.

Bolstad et al. [4] провели анализ методов нормализации и сравнили их на двух разных выборках микрочипов. По результатам их исследования методы, не использующие эталон, позволяют уменьшить вариацию интенсивностей проб из одного подмножества лучше, чем другие методы нормализации. При попарной нормализации наименьшее различие интенсивностей микрочипов получалось при квантильной нормализации. Применение методов, не использующих эталон, приводят к схожим результатам, однако метод квантильной нормализации работает несколько лучше. Кроме того, это самый быстрый метод нормализации. Квантильная нормализация входит в комплекс предобработки данных RMA.

2.1.3 Суммаризация

Суммаризацией называется процесс вычисления значения экспрессии гена по интенсивностям свечения (возможно, преобразованным) соответствующих этому гену проб. Для вычисления конечной оценки экпрессии были предложены несколько методов [2],[21]. Самый простой метод суммаризации представляет значение экспрессии

гена в виде среднего значения интенсивностей соответствующих ему проб:

$$C_j = \frac{1}{m} \sum_{i=1}^{m} I_i^j,$$

где m — число соответствующих гену j проб. Наиболее популярными являются следующие методы.

- **Tukey Bi-Weight**

  Алгоритм, основанный на Tukey Bi-Weight, входит в комплекс обработки данных MAS 5.0. Значение экспрессии гена j вычисляется как робастное среднее предобработанных нормализованных интенсивностей проб, соответствующих гену j. Алгоритм состоит из следующих шагов [2]:

  1. Вычисляется среднее M^j для набора интенсивностей проб $I_1^j, \ldots, I_m^j$, соответствующих гену j.
  2. Вычисляется медиана S множества $\{|I_i^j - M^j| \;\big|\; i = 1, \ldots, m\}$.
  3. Для каждого значения интенсивности вычисляется расстояние до центра: $u_i = \frac{I_i^j - M^j}{cS + \epsilon}$, $i = 1, \ldots, m$, где c — константа со значением по умолчанию 5, $\epsilon > 0$ предотвращает деление на ноль.
  4. Вычисленные расстояния используются для расчёта весов интенсивностей проб:

     $$\omega(u) = \begin{cases} (1 - u^2)^2, |u| \leqslant 1 \\ 0, |u| > 1 \end{cases}$$

     Чем больше расстояние от значения интенсивности до центра, тем меньший вес получает проба. Вес проб-выбросов может принимать значение 0.
  5. Вычисляется взвешенное среднее Тьюки

     $$T_j = \frac{\sum_{i=1}^{m} \omega(u) I_i^j}{\sum_{i=1}^{m} \omega(u)}$$

     Значение T_j называется величиной экспрессии гена j.

  Отметим, что алгоритм суммаризации учитывает интенсивности проб каждого чипа по-отдельности.

- **Median Polish**

  Суммаризация с помощью Tukey Bi-Weight учитывает информацию только об одном чипе. Хотя значения интенсивностей имеют разные веса, алгоритм рассматривает все пробы одинаково. Однако Li et Wong [21] показали, что пробы по-разному взаимодействуют со специфическими генами. Модель, на которой базируется метод MedianPolish, учитывает различия в характере взаимодействия проб с генами. MedianPolish входит в комплекс обработки данных RMA [18].

  Интенсивность i-й пробы на t-м чипе, соответствующая j-му гену, представляется в виде

  $$\log_2(I_{it}^j) = \mu_t^j + \alpha_{it}^j + \epsilon_{it}^j,$$

  где α_{it}^j — коэффициент взаимодействия i-пробы с генами, $\sum_i \alpha_{it}^j = 1 \; \forall t$, μ_t^j — концентрация j-го гена на t-м чипе, ϵ_{it}^j — шумовая константа с нулевым математическим ожиданием. Оценка $\hat{\mu}_t^j$ будет искомым значением концентрации j-го гена на t-м чипе.

  Чтобы оценить концентрацию μ_t^j применяется алгоритм MedianPolish, дающий робастную оценку по двум причинам. Во-первых, вычисление медианы вместо среднего позволяет исключить влияние выбросов. Во-вторых, этот алгоритм учитывает значение интенсивностей на всех чипах выборки. Модель, настраиваемая алгоритмом MedianPolish, выглядит следующим образом: $I_{it} = \mu + \alpha_i + \beta_t + \epsilon_{it}$, где I_{it} — это матрица интенсивностей проб, соответствующих одному гену. Матрица оценок ϵ_{it} вычисляется путём поочерёдного вычитания медиан столбцов и строк из элементов матрицы I_{it}. На каждой итерации к значениям вектора α и вектора β прибавляются медиана строк и медиана столбцов соответственно. Этот процесс повторяется до тех пор, пока изменение матрицы не будет достаточно мало, либо пока не будет достигнуто максимальное число итераций. Оценка $\hat{\mu}$ определяется как сумма медиан векторов α и β, а оценки $\hat{\alpha}_i$ и $\hat{\beta}_j$ равны $a_i - \hat{\mu}$ и $b_j - \hat{\mu}$ соответственно. Окончательная оценка уровня экспрессии будет равна $\hat{\mu}_j = \hat{\mu} + \hat{\beta}_j$.

- **FARMS (Factor Analysis For Robust Multiarray Summarization)**
  Hochreiter et al. [17] предложили метод суммаризации, базирующийся на предположении о нормальности распределения логарифма интенсивности свечения проб:

$$\log(I) = \lambda z + \epsilon, \qquad I, \lambda \in \mathbb{R}^n, \;\; z \sim \mathcal{N}(0,1), \;\; \epsilon \sim \mathcal{N}(0, \Psi).$$

  Здесь $\mathcal{N}(0, \Psi)$ — многомерное нормальное распределение с нулевым матожиданием и матрицей ковариации Ψ, z называется *фактором*, z и ϵ статистически независимы.

  Согласно этой модели, логарифм интенсивности распределён нормально по закону $\log(I) \sim \mathcal{N}(0, \lambda\lambda^T + \Psi)$. Алгоритм определения истинного сигнала состоит из трёх шагов:

  1. Нормализация наблюдаемого сигнала для выполнения предположений модели. Логарифмы интенсивности нормализуются таким образом, чтобы математическое ожидание было нулевым.
  2. Максимальный апостериорный факторный анализ. Используется метод максимального правдоподобия оценки параметров апостериорных плотностей вероятности $p(\lambda, \Psi | \{\log(I)\})$, $p(\{\log(I)\} | \lambda, \Psi)$.
  3. Определение истинного сигнала на основании оцененных параметров.

  Авторы утверждают, что их алгоритм лучше работает на зашумлённых данных, чем другие методы.

2.2 Биологическая вариация

На поверхности микрочипа происходят сложные процессы, описать и предсказать которые можно лишь с определённой точностью. При изготовлении микрочипов Affymetrix пробы были разработаны таким образом, чтобы они были полностью комплементарны уникальному участку определённого гена и при этом некомплементарны другим генам. Однако на практике оказывается, что достаточно часто для проб существуют гены, последовательность которых частично комплементарна последовательности пробы. В таком случае проба тоже могут вступить в реакцию с

геном, хотя и с меньшей вероятностью, чем со специфическим геном. Эта реакция называется *кросс-гибридизацией* и вносит существенный вклад в интенсивность свечения проб.

К тому же на протекание реакции и силу связывания молекул оказывает влияние их длина и нуклеотидный состав. Известно, что молекулы, содержащие больше гуанина (G) и цитозина (C), обладают большей силой гибридизации [20]. Некоторые авторы рассматривают влияние неполного синтеза проб на силу образующихся связей с генами [26].

Многие исследователи [5][13][26] считают важным рассмотрение этапа промывки микрочипа, когда с поверхности удаляется образец с негибридизованными молекулами генов. На этом этапе происходит распад некоторых соединений проб с генами. Более того, при больших концентрациях генов зависимость свечения проб от концентрации перестаёт быть линейной [6]. Это происходит в силу насыщения: большая часть молекул проб уже вступила в реакцию с молекулами генов, и при увеличении концентраций генов новых связей образуется всё меньше.

Мы остановимся на двух важных эффектах, возникающих при протекании реакции на микрочипе: кросс-гибридизации и насыщении.

2.2.1 Эффект кросс-гибридизации

Одной из задач при анализе данных микрочипов ДНК является распознавание ложного сигнала, возникающего вследствие неспецифических взаимодействий между пробами и молекулами генов. Этот эффект неизбежно возникает в ходе реакции сложной смеси фрагментов ДНК/РНК с миллионами проб на микрочипе одновременно [11]. Разработчики платформы Affymetrix создали пары проб: PM (Perfect Match) и MM (MisMatch). PM-пробы полностью комплементарны участкам специфичных генов, а MM-пробы по составу идентичны PM за исключением одного нуклеотида в центре (13-го). Предполагалось, что интенсивности сигнала MM-проб позволят оценить неспецифические взаимодействия соответствующих PM-проб и таким образом выделить чистый сигнал PM-пробы. Однако у этого метода существует две проблемы. Во-первых, оказалось, что 30% интенсивностей MM-проб больше, чем интенсивности соответствующих PM-проб и оценить кросс-гибридизацию для таких

пар невозможно. Во-вторых, необходимо нанести на микрочип большое количество MM-проб, равное количеству PM-проб. Таким образом, уменьшается количество PM-проб, которое возможно нанести на чип, соответственно меньше генов можно выявить с помощью микрочипа.

На более новых микрочипах Affymetrix MM-пробы уже отсутсвуют, вместо них появились так называемые пробы шума. Эти пробы сконструированы таким образом, чтобы для них не было специфических генов, поэтому их интенсивность может быть мерой неспецифических взаимодействий. Пробы шума отличаются количеством G и C нуклеотидов (гуанина и цитозина). Сигнал остальных проб, вызванный неспецифическими взаимодействиями, оценивается как медиана сигнала проб шума, имеющих в составе такое же количество GC-нуклеотидов.

Так как при использовании малого количества проб шума появляется возможность нанести на чип больше PM-проб и таким образом повысить точность микрочипов, разработка и улучшение методов оценки кросс-гибридизации является важной и перспективной темой в анализе данных микрочипов ДНК.

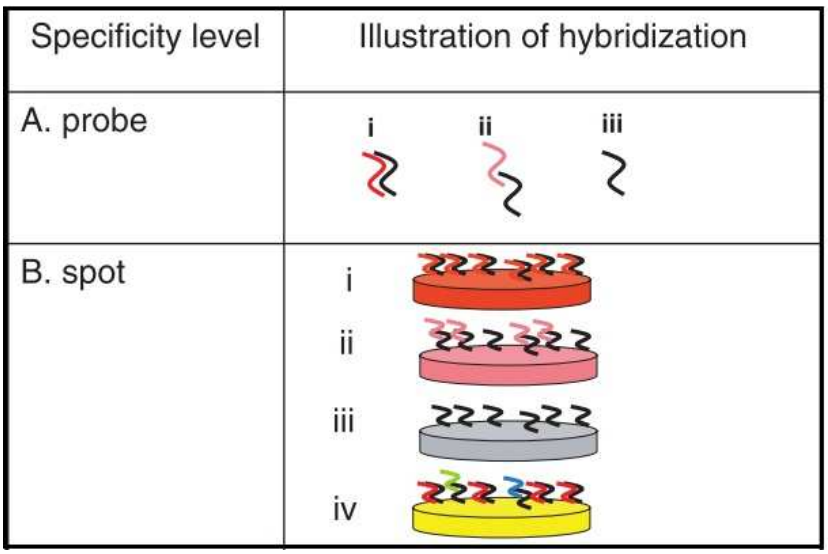


Рис. 5: Эффект кросс-гибридизации на поверхности микрочипа. На уровне проб (A.probe) изображены 3 варианта гибридизации: полная (i), частичная (ii) и отсутствие (iii). На уровне набора проб (B.spot) могут быть 4 варианта: все пробы связаны со специфическими генами (i), часть проб связана с неспецифическими генами (ii), пробы не связаны (iii), возможны все типы взаимодействий (iv)

2.2.2 Эффект насыщения

Проведенные исследования показали, что зависимость между интенсивностью свечения проб и концентрацией генов не линейна [6] (Рис. 6). Интенсивность, измеряемая при сканировании, это мера свечения флюоресцентной метки, прикрепленной к месту соединения молекулы гена с пробой.

Процесс гибридизации на поверхости микрочипа можно схематично описать следующим образом [15]:

Проба + Ген $\underset{k_b}{\overset{k_f}{\rightleftarrows}}$ Соединение пробы и гена (двойная спираль),

где k_f и k_b — параметры реакции образования и распада соединений проб с молекулами генов. Отметим, что реакция может протекать в двух направлениях: образование связей проба-ген и распад таких связей.

Пусть n_p — число молекул пробы, n_0 — число молекул специфического для данной пробы гена в образце, n_B — количество образовавшихся соединений молекул пробы с молекулами гена, V_{probe} — объём, занимаемый пробой, V_{total} — общий объём образца. Тогда концентрации молекул проб, свободных молекул генов и образовавшихся соединений представляются в виде

$$[P] = \frac{n_p - n_B}{V_{probe}}, \qquad [G] = \frac{n_0 - n_B}{V_{total}}, \qquad [PG] = \frac{n_B}{V_{total}}.$$

Количество образующихся соединений в момент времени выражается уравнением

$$\frac{\partial n_B}{\partial t} = k_f(n_p - n_B)\Big(\frac{n_0 - n_B}{V_{total}}\Big) - k_b n_B.$$

Предполагая, что в системе достигнуто равновесие, то есть $\frac{\partial n_B}{\partial t} = 0$, и что $n_p \ll n_0$, получаем уравнение

$$n_B = \frac{n_p n_0}{\frac{k_B}{k_f} V_{total} + n_0}.$$

Переходя от молекул к интенсивностям проб и исходным концентрациям, получаем уравнение

$$I = \frac{Ac}{K + c} + I_0.$$

Здесь $c = \frac{n_0/V_{total}}{N_A/liter}$ — концентрация в моль/литр ($N_A = 6.02 \times 10^{23}$ моль$^{-1}$ — число Авогадро); $K = \frac{k_B/k_f}{N_A/liter}$, а A отличается от n_p пропорционально тому, как n_B отличается от интенсивности пробы I. Константа I_0 была введена для того, чтобы учесть неспецифические взаимодействия на поверхности микрочипа.

Эта модель, называемая моделью адсорбции Ленгмюра, легла в основу большинства физико-химических моделей данных микрочипов ДНК. Подробнее о модификациях этой модели будет описано ниже.

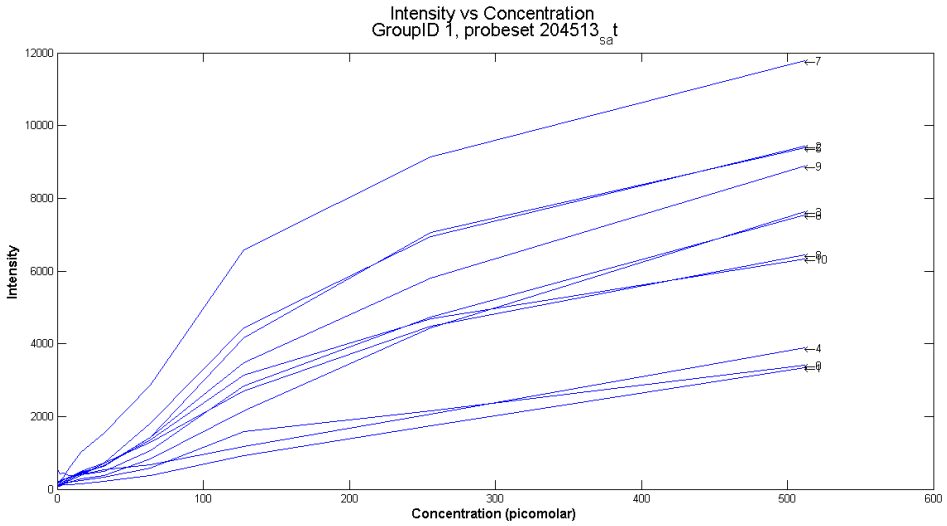


Рис. 6: График зависимости интенсивности свечения проб, соответствующих одному гену, от концентрации специфичного гена. Можно заметить, что кривая зависимости для некоторых проб может быть аппроксимирована линейной функцией, а для некоторых проб, очевидно, имеет нелинейный характер.

3 Определение экспрессии генов с помощью микрочипов ДНК

3.1 Эвристические подходы

Распространенные подходы к оценке экспрессии генов, как правило, состоят из трёх этапов: вычитания фонового шума, нормализации данных и суммаризации интенсивностей соответствующих генам проб. Многие методы абсолютно эмпирические: не принимаются во внимание физические процессы, протекающие на этапе гибридизации, эффекты насыщения при больших концентрациях и специфичные особенности проб. Схемы работы нескольких наиболее популярных методов отражены в таблице 3.1.

Метод	Фоновая поправка	Нормализация	Суммаризация
MAS 5.0	MAS 5.0	Стандартный линейный метод	Tukey Biweight
RMA	RMA	Квантильная нормализация	MedianPolish
DFCM	DFCM	Квантильная нормализация	Tukey Biweight
FARMS	—	Квантильная нормализация	FARMS

Эти методы ориентированы на определение изменения уровня экспрессии одного гена в пределах одного эксперимента. Они не дают значений исходных концентраций генов, результат их работы является величной отношения уровней экспрессии. Однако в тех случаях, когда необходимо сравнить уровни экспрессии разных генов или одного и того же гена в разных экспериментах, эти методы оказываются неприменимы.

Более того, Burden et al. [6] утверждают, что изменение уровней экспрессии связаны с изменениями интенсивностей не линейно, как обычно предполагается.

Hekstra et al. [14] подчёркивают, что существующие методы направлены на измерение изменения уровня экспрессии одного гена, но не позволяют сравнить экспрессию разных генов. Поэтому были разработаны методы, принимающие во внимание процессы, протекающие на поверхности чипа, индивидуальные особенности проб и генов.

3.2 Физико-химические модели

С развитием методов обработки данных микрочипов ДНК появились модели, основанные на принципах физической химии [14],[15]. Физико-химические модели гибридизации на микрочипе учитывают взаимодействия проб со специфическими и неспецифическими генами, связывание свободных молекул в образце, эффекты насыщения, обеднения смеси и этап промывки чипа. Эти модели, базирующиеся в основном на химической модели адсорбции Ленгмюра, позволяют предсказать абсолютные концентрации генов.

В отличие от методов, ориентированных на оценивание изменения уровня экспрессии генов, эти модели потенциально подходят для сравнения уровней экспрессии разных генов в одном или разных экспериментах.

Модель адсорбции Ленгмюра взята за основу во многих методах обработки данных микрочипов ДНК. Модель учитывает эффект насыщения, заключающийся в том, что при избытке молекул РНК в образце количество гибридизованных проб практически не меняется и интенсивность свечения проб не увеличивается. Общий вид интенсивности в модели Ленгмюра:

$$I_i = \frac{Ac}{c + K} + I_0.$$

Здесь I_i — интенсивность i-й пробы, c — концентрация соответствующего пробе гена. Сумма $I_0 + A$ — это интенсивность при насыщении, то есть при максимальной концентрации специфических генов. I_0 — это фоновая компонента, в которой учтены технические погрешности при сканировании и биологические вариации (например, кросс-гибридизация). K — это концентрация специфического гена, соответствующая «полунасыщению». Этот коэффициент возникает из термодинамических соображений, он пропорционален свободной энергии связывания проб и участков специфических генов.

В работе Mulders et al. [24] значение A берётся постоянным, например, равным среднему значению максимальных интенсивностей проб на нескольких микрочипах. Коэффициент K может учитывать различные факторы гибридизации, такие как эффективность пробы, энергию гибридизации пробы, связывание свободных участков РНК друг с другом, кросс-гибридизацию проб. В [24] K имеет вид

$$K = \frac{exp(\Delta G/RT)}{1 + \tilde{c}K_0 exp(\Delta G'/RT')}$$

Здесь ΔG и $\Delta G'$ — энергии гибридизации проб и генов, зависящие от последовательности нуклеотидов, R — универсальная газовая постоянная (1.99 кал/(моль·K)), K_0 — константа. Параметры $T, T', \Delta G', K_0$ настраиваются на данных с известными концентрациями генов (специальных экспериментах) с помощью восстановления регрессии. Heim et al. [13] показали, что параметры, оцененные с помощью данных одного эксперимента, подходят для определения экспрессии генов в других экспериментах, что подтверждает возможность построения универсальной модели для микрочипов ДНК.

Burden et al. [6] разработали несколько динамических моделей адсорбции — модификаций модели Ленгмюра, связывающих интенсивность свечения проб и концен-

трации соответствующих участков РНК. Авторы сравнивают равновесные и неравновесные модели с разными параметрами на данных эксперимента «Латинский Квадрат» (см. Рис. 8). В статье было показано, что для этих данных наиболее подходящей является равновесная модель Ленгмюра с параметрами I_0, A, K, зависящими от соответствующих пробы и гена. В этом случае информация о составе проб может быть использована для оценки параметров модели для улучшения оценки абсолютной концентрации генов.

Метод оценки параметров, предложенный Held et al. [15], оказался не самым оптимальным. В этом случае параметры A, K оценивались с помощью свободной энергии гибридизации, однако параметр I_0 был постоянным для всех генов. Это противоречит результатам Burden et al. [6], согласно которым все три параметра должны зависеть как от пробы, так и от гена.

Для оценки параметров A, K, I_0 Hekstra et al. [14] использовали линейную регрессию. При этом учитывался состав проб и способность проб вступать в специфические реакции. Кроме того, авторы пришли к выводу, что оценка технического шума отдельно от модели увеличивает её чувствительность. Таким образом, они вводят параметр ϵ технического шума и оценивают его отдельно от I_0 до применения линейной регрессии. Вопрос определения конечных концентраций заслуживает отдельного внимания. В Hekstra et al. [14] логарифм концентрации гена представляется в виде суммы логарифмов оцененных концентраций проб, причём «дефектные» пробы не учитываются. Под «дефектными» авторы подразумевают пробы, концентрация соответствующих генов для которых оказалась отрицательной. Логарифм концентрации представляется в виде:

$$\ln(\hat{c}_{gene}) = \frac{1}{n_S} \sum_{p \in S} \ln(\hat{c}_p).$$

Здесь S — это подмножество проб, соответствующих данному гену, n_S — число элементов в этом подмножестве. Кроме среднего авторы пробовали применять и другие функции, такие как медиана и M-оценки (обощения оценки максимального правдоподобия), однако они не принесли существенных изменений.

Burden et al. [6] предлагают не удалять «дефектные пробы», потому что таким образом слишком много проб приходится не учитывать, а заменять значение концен-

трации гена для таких проб на некоторую константу. Кроме того, вместо среднего было предложено использовать медиану: $\ln(\hat{c}_{gene}) = median(\ln(\hat{c}_p))$.

Кроме модификаций модели Ленгмюра, Burden et al. рассмотрели изотерму Сипса, описывающую зависимость интенсивности свечения проб от концентрации генов в предположении о нормальном распределении свободной энергии. Общий вид модели:

$$I_i = \frac{Ac^\alpha}{c^\alpha + K^\alpha} + I_0.$$

Здесь параметр α принимает значения от 0 до 1, с увеличением α уменьшается вариация распределения. Модель Сипса может быть уместна при больших концентрациях генов, однако в большинстве случаев модель с параметром $\alpha = 1$ (то есть стандартная модель Ленгмюра) оказывается оптимальной.

Ono et al. [27] сравнивают модель Ленгмюра с моделью, предложенной Zhang et al. [30], и предлагают свою модель конечной гибридизации (Finite Hybridization Model). Модель Ленгмюра рассматривается в немного модифицированном варианте:

$$I^{Langmoir} = \alpha\frac{Kc}{1 + Kc} + I_0,$$

где α — константа нормализации, K — константа равновесия образования связей проб с генами. Константа K здесь, как и в предыдущей модели, пропорциональна свободной энергии связывания проб и генов и зависит от нуклеотидного состава пробы.

Модель, предложенная Zhang et al. [30], отличается тем, что в явном виде учитывает неспецифические взаимодействия. Интенсивность в ней имеет вид

$$I^{Zhang} = \alpha\left(\frac{c}{1 + K'} + \frac{N}{1 + K''}\right) + I_0,$$

где c — концентрация специфического гена, N — концентрации неспецифических генов, с которыми может возникнуть кросс-гибридизация. K' и K'' — константы равновесия образования связей, пропорциональные свободной энергии специфических и неспецифических взаимодействий соответственно. В этой модели также предполагается, что с увеличением константы взаимодействия пробы с генами K', K'' ускоряется насыщение интенсивности, однако уровень насыщения прямо пропорционален концентрации специфических генов. Модель описывает состояние, в котором все специфические гены связаны с соответствующими пробами.

Модель конечной гибридизации (Finite Hybridization Model), предложенная Ono et al. [27], имеет вид

$$I^{FH} = N(d^{sp}[PT^{sp}] + d^{ns}[PT^{ns}]) + I_0.$$

Здесь $[PT^{sp}]$ и $[PT^{ns}]$ — число образовавшихся связанных молекул со специфическими и неспецифическими генами соответственно, d^{sp} и d^{ns} — коэффициенты диссоциации (распада связей проба-ген) специфических и неспецифических взаимодействий, N — константа нормализации. $[PT^{sp}]$ и $[PT^{ns}]$, в свою очередь, зависят от коэффициентов взаимодействия проб со специфическими и неспецифическими генами K' и K'', количеством свободных проб и свободных молекул генов в образце.

Диссоциация связей проб с генами на этапе удаления остатков образца с микрочипа рассматривалась Burden [7], Held [16] как неравновесный процесс, который ухудшает флюоресцентный сигнал. Чтобы учесть этот эффект, Ono et al. ввели коэффициенты d^{sp} и d^{ns}.

Таким образом, модель конечной гибридизации позволяет учесть больше факторов, влияющих на интенсивность свечения проб, таких как эффект насыщения, диссоциация связей, кросс-гибридизация, разная длина проб и т.д. Авторы утверждают, что с помощью предложенной ими модели можно точнее предсказать концентрации генов эксперимента «Латинский квадрат» по сравнению с обычной моделью Ленгмюра и моделью Zhang.

Burden и Binder [5] также рассматривают физико-химические модели, учитывающие диссоциацию связей проб с генами. Помимо учёта гибридизации, связывания свободных молекул в растворе и сворачивания проб и генов, важным новшеством является детальное рассмотрение процесса обеднения смеси (depletion), то есть уменьшение числа молекул генов вследствие гибридизации. Учитывать уменьшение концентрации генов имеет смысл в том случае, когда концентрация генов сравнима с концентрацией молекул проб (для эксперимента «Латинский Квадрат» авторы приводят значение 200 пМ.). Таким образом, эффект обеднения необходимо учитывать при оценке маленьких концентраций генов.

Burden и Binder рассматривают локальную и глобальную модели обеднения смеси. В первом случае (фактически, модель Ono et al. [27]) уменьшение молекул специфического гена вследствие гибридизации с конкретной пробой влияет на интен-

сивность только этой пробы. Эта модель имеет место в том случае, когда пробы, соответствующие одному гену, расположены далеко друг от друга на поверхности микрочипа. Во второй модели, предложенной Burden и Binder, учитывается глобальное обеднение смеси. В этой модели интенсивности нескольких или всех проб, соответствующих одному гену, зависят от уменьшения числа молекул гена вследствие гибридизации. Этот эффект можно наблюдать на микрочипах, на которых предусмотрены PM и MM-пробы (в том числе, в эксперименте «Латинский квадрат»).

О зависимости между интенсивностями PM и MM-проб и неспособности базовой модели Ленгмюра описать различия между ними пишут также Burden, Pittelkow и Wilson [7]. Авторы замечают, что асимптотически интенсивность свечения MM-пробы при больших концентрациях специфического гена должна быть меньше, чем интенсивность PM-пробы, однако модель Ленгмюра этого не учитывает. Интенсивность MM-пробы зависит не только от кросс-гибридизации и шумовой компоненты [7],[9],[18],[26]. При высоких концентрациях генов большая часть MM-проб гибридизуется также со специфическими генами. Поэтому авторы рассматривают MM-пробы как менее чувствительные модификации PM-проб и исследуют различия в интенсивности свечения проб при концентрациях специфических генов, при которых возникает насыщение.

При построении модели Burden et al. фокусируются на гибридизации на поверхности микрочипа и последующем этапе промывки чипа, на котором смываются негибридизованные молекулы в образце. Авторы рассматривают модели Hekstra [14] и Halperin [12], учитывающие неспецифичную гибридизацию. Halperin et al. учитывают в модели гибридизации параметр электростатической свободной энергии. В этом случае кривая зависимости интенсивности свечения от концентрации может отклоняться от гиперболы, но разницу в насыщении MM и PM-проб в рамках этой модели объяснить не удаётся. Авторы рассмотрели модель, предполагающую, что равновесие в ходе реакции гибридизации не было достигнуто. В этой модели появляется дополнительная степень свободы, однако статистические тесты показали, что этот параметр не является значимым. Кроме того, авторы рассмотрели изотерму Сипса, параметр α которой согласно статистическим критериям проверки значимости так-

же оказался незначимым. Это совпадает с результатами, полученными в [6]. Поэтому авторы предлагают использовать модификации равновесной модели Ленгмюра.

Рассматривались также модели, учитывающие сворачивание молекул генов и проб, связывание проб друг с другом и гибридизацию свободных молекул генов в образце. Все эти факторы оказались незначительными и неспособными объяснить различие интенсивностей при насыщении PM и MM-проб. Однако авторы считают, что объяснить этот факт можно, если учитывать этап промывки микрочипа.

В работе Nguyen [26] рассматривается влияние возможного неполного синтеза проб на процесс гибридизации. Автор учитывает тот факт, что при нанесении проб на поверхность микрочипа реакция не может протекать с точностью 100%, и некоторые пробы могут иметь отличную от 25 нуклеотидов длину. Это может повлиять на силу взаимодействия пробы с молекулами генов. В этой модели коэффициент взаимодействия пробы и гена K зависит от свободной энергии гибридизации, которая, в свою очередь, зависит не только от нуклеотидного состава пробы, но и от её длины. Однако эксперименты показали, что учёт различной длины проб не оказывает большого эффекта на качество восстанавливаемой зависимости.

3.3 Модели кросс-гибридизации

Кросс-гибридизация неизбежно возникает в ходе протекания реакций на поверхности микрочипа, поэтому исследователи уделяют пристальное внимание разработке методов, позволяющих учитывать этот эффект в анализе данных. Кроме методов, предложенных разработчиками микрочипов (использования MM-проб или специальных фоновых проб), были разработаны подходы, зависящие от конструкции конкретного вида микрочипов.

Так, Naef и Magnasco [25] рассматривают биофизическую природу взаимодействий на чипе и объясняют тот факт, что MM-пробы светятся лучше, чем PM-пробы. На это влияют связи между олигонуклеотидом и флюоресцентной меткой. Авторы также представляют логарифм светимости в виде суммы нескольких переменных, значение которых зависит от того, какой нуклеотид находится в данном месте. Модель ближайшего соседа предложили Zhang et al. [30]. Авторы тоже объясняют «феномен» светимости MM-проб с физической точки зрения. В их модели интенсивность

раскладывается на три компоненты: шум, специфические и неспецифические взаимодействия молекул. В формуле фигурирует экспонента энергии, а энергия как раз зависит от состава пробы и положения нуклеотидов в ней. В качестве достоинства модели авторы отмечают, что с её помощью удалось уменьшить вариацию интенсивностей.

Furusawa et al. [11] оценивали кросс-гибридизацию, опираясь на предположение о том, что только ограниченное количество генов может вступать с пробами в неспецифические взаимодействия. В своей модели авторы стремились учесть влияние вторичной структуры молекул на формирование неспецифических связей.

Wu et al. [29] предлагают моделировать кросс-гибридизацию с помощью свободной энергии. Их исследования показали, что кросс-гибридизация чаще всего происходит при пересечении участков пробы и гена на 10-16 нуклеотидов. Авторы также отмечают, что с помощью модели кросс-гибридизации можно выявлять склонные к неспецифическим взаимодействиям пробы, чтобы учитывать их в дальнейшем при разработке микрочипов. Wu et al. рассматривают линейную модель зависимости интенсивности от концентрации, потому что для проб, вступающих в неспецифические реакции, не выполняется закон насыщения.

3.3.1 Оценка кросс-гибридизации с помощью BLASTN

Существует несколько методов для оценки склонности к взаимодействию между молекулами, состоящими из нуклеотидов. Однако для того, чтобы определить коэффициенты взаимодействия всех проб со всеми генами, большинству из этих методов понадобится очень много времени и вычислительных мощностей. Поэтому распространённым решением является поиск возможных взаимодействий на основании алгоритма BLASTN выравнивания символьных последовательностей.

Если выравнивание найдено, коэффициент взаимодействия может быть вычислен на основании длины совпадающих участков, процента совпадения, значения выравнивания или величины E(Expect). E-величина вычисляется алгоритмом BLASTN и отражает число таких же выравниваний, которые могут оказаться в случайной выборке такого же размера. Чем меньше E-величина, тем более значимым считается найденное выравнивание.

Финальное значение выравнивания, определяемое алгоритмом BLASTN, вычисляется с учётом длины пересекающихся участков, количества вставок и удалений (делеций) нуклеотидов, а также типа совпавших нуклеотидов. Тип нуклеотидов учитывается с помощью матрицы 4×4, в которой каждое значение соответствует возможной паре нуклеотидов. Значение выравнивания есть сумма соответствующих элементов матрицы за вычетом штрафов за вставки и делеции.

Вычислительная эффективность BLASTN делает его привлекательным для определения коэффициентов взаимодействия проб и генов, поэтому Eklunf et al. [10] решили из эмпирических соображений модифицировать коэффициенты матрицы нуклеотидов и величину штрафов за пропуски. Для этого они использовали данные эксперимента «Латинский квадрат» и подбирали параметры так, чтобы коэффициенты взаимодействия коррелировали с изменением интенсивности свечения проб. Оптимизация параметров алгоритма BLASTN может увеличить точность выявления потенциально склонных к кросс-гибридизации проб, что может быть полезным для определения истинных источников изменения интенсивности свечения проб и в разработке новых микрочипов.

Тем не менее, выравнивание предоставляет лишь грубую оценку склонности проб к взаимодействию с генами. Для более точного определения коэффициентов взаимодействия необходимо учитывать энергию взаимодействия нуклеотидов, позиции нуклеотидов в молекуле, сложность последовательности и другие факторы, которые не могут быть учтены с помощью алгоритма BLASTN.

Часть II

Модель данных микрочипов ДНК, учитывающая эффекты кросс-гибридизации и насыщения

3.4 Мотивация данной работы

Несмотря на обилие существующих методов анализа данных микрочипов ДНК, исследования в этой области не теряют своей актуальности. В данной работе мы попытались учесть достоинства и недостатки существующих методов и построить модель, включающую в себя наиболее значимые эффекты, возникающие при гибридизации.

Во многих работах отмечалось, что при реакции исследуемого образца с пробами на поверхности микрочипа неизбежно возникает кросс-гибридизация [8], [11]. Исследования Eklund et al. [10] показали, что зависимость интенсивности свечения от концентрации гораздо больше проявляется на уровне отдельных проб, чем на уровне набора проб, соответствующих одному гену. Wu et al. [29] также отмечали, что кросс-гибридизация характерна для индивидуальных проб, так как она зависит от особенностей строения каждой пробы. О бесполезности MM-проб как измерителей кросс-гибридизации упоминалось достаточно, однако Wu et al. показали, что для эксперимента «Латинский квадрат» PM и MM-пробы имеют схожее поведение. Это означает, что MM-пробы тоже могут вступать в реакцию со специфичными генами и их можно учитывать при определении исходной концентрации.

Характер зависимости интенсивности свечения от концентрации генов был предметом обсуждений для многих исследователей [1], [14], [15], [24]. Необходимо учитывать, что при высоких концентрациях специфичных генов зависимость интенсивности от концентрации перестаёт быть линейной, однако для MM-проб это не выполняется, так как сила связей меньше и насыщения не возникает. Мы попытались

построить гибкую модель, позволяющую учесть как линейный, так и нелинейный характер зависимости интенсивности от концентрации.

Эффекты неполного синтеза проб, реакций связывания свободных молекул в образце друг с другом, процесса промывки чипа и сворачивания проб и генов оказались незначимыми по результатам исследований [5], [26], поэтому не будут учитываться в предлагаемой модели.

В данной работе предлагается модель данных микрочипов ДНК, учитывающая эффекты кросс-гибридизации и насыщения. Зависимость между интенсивностью свечения и концентрацией генов в образце строится на уровне отдельных проб с учётом индивидуальных особенностей их строения. Кроме того, эффект кросс-гибридизации предлагается учитывать не только при определении сигнала, соответствующего специфическим взаимодействиям, но и при суммаризации. Важным новшеством на этапе определения исходной концентрации гена является учёт интенсивностей свечения всех проб, которые могут вступать как в специфические, так и в неспецифические взаимодействия с данным геном.

3.5 Модель кросс-гибридизации

В данной работе предлагается модель данных микрочипа ДНК, учитывающая кросс-гибридизацию проб и технический шум. Как уже отмечалось ранее, пробы на поверхности чипа могут вступать в реакцию с молекулами специфических генов, которым они полностью комплементарны, и с некоторыми другими генами, которым они комплементарны лишь частично. Чтобы описать эти взаимодействия, введём матрицу $A = \{A_{ij}\}$, которую назовём матрицей взаимодействия проб и генов. Каждый коэффициент $A_{ij} \geqslant 0$ матрицы A выражает склонность i-й пробы к взаимодействию с j-м геном. Матрица A имеет размерность $P \times N$, где P — количество проб на микрочипе, N — число генов, состав которых известен и взаимодействие с которыми предусмотрено разработчиками микрочипа. Микрочипы Affymetrix ориентированы на распознавание более 22 000 генов и содержат от 500 000 до 900 000 проб. Каждая проба соответствует только одному специфическому гену и, возможно, может вступать в неспецифические взаимодействия с несколькими другими генами. Таким образом, матрица A является сильно разреженной матрицей большой размерности.

В предлагаемой модели интенсивность свечения пробы выражается через концентрации всех генов, которые могут быть в образце. Зависимость имеет вид:

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

где t — номер чипа, i — номер пробы; суммирование идёт по всем генам, которые могут присутствовать в образце, $A_{ij} \geqslant 0$ — коэффициент склонности i-й пробы к взаимодействию с j-м геном, $C_j^t \geqslant 0$ — концентрация j-го гена на t-м чипе, d^t — параметр нормализации, b_i^t — фоновая поправка.

3.6 Восстановление параметров модели

Параметры d_t были вычислены с помощью алгоритма квантильной нормализации. Этот метод был выбран лучшим согласно Bolstad et al.[4]. Для вычисления фоновой поправки b_i^t был использован алгоритм MedianPolish, учитывающий информацию обо всех микрочипах в эксперименте, тем самым позволяющий учесть индивидуальные характеристики проб.

Основной задачей являлось восстановление коэффициентов матрицы взаимодействий A. В предположении, что концентрации генов C_j^t заданы или фиксированы, предлагается искать коэффициенты A_{ij} путём минимизации функционала квадратичной невязки:

$$Q = \sum_t \left(\sum_i (I_i^t - Y_i^t)^2 + \lambda \sum_{ij} |A_{ij}| \right) \to \min_{A_{ij}},$$

где Y_i^t — наблюдаемая интенсивность, λ — параметр регуляризации.

L_1-регуляризация была введена из тех соображений, что матрица A должна быть сильно разрежена, потому что большая часть коэффициентов взаимодействия проб с генами равна нулю. Варьируя параметр регуляризации λ можно регулировать число нулевых коэффициентов взаимодействия для каждой пробы. Кроме того, задав в качестве параметра регуляризации вектор $\vec{\lambda} = \{\lambda_{ij}\}$, можно, например, поощрять коэффициенты специфического взаимодействия. Подробнее эксперименты с различными параметрами регуляризации будут описаны в разделе Эксперименты.

3.6.1 Алгоритм выравнивания последовательностей BLAST

Для того, чтобы оценить правильность восстановления коэффициентов матрицы A взаимодействий проб с генами, было предложено восстановить матрицу A^{blast} с помощью алгоритма выравнивания символьных последовательностей BLASTN. Алгоритм BLASTN — специальная модификация алгоритма BLAST [3] для быстрого сравнения последовательностей нуклеотидов. Мы использовали BLASTN для определения длины участка пересечения последовательностей пробы и гена. Алгоритм основан на выравнивании последовательности пробы и перевернутой последовательности, комплементарной к гену. Для построения матрицы A^{blast} на вход алгоритму BLASTN были поданы последовательности проб [2] и последовательности генов из сборки человеческого генома версии 18, использованной при создании чипов Affymetrix. Отметим, что последовательности проб и генов зависят от модели микрочипа. Пример работы алгоритма BLASTN изображён на Рис.(7).

Коэффициенты матрицы A_{ij}^{blast} представляет собой число совпадающих нуклеотидов в последовательности i-й пробы и j-го гена. В идею использования матрицы A^{blast} для оценки адекватности предлагаемой модели заложено предположение о том, что некомплементарные молекулы не могут вступать в реакцию. Таким образом, если коэффициент матрицы A_{ij}^{blast} равен нулю, то проба и ген некомплементарны, и коэффициент матрицы A_{ij} тоже должен быть равен нулю.

Похожую матрицу взаимодействий восстанавливали Wu et al. [29]. На основании своих исследований авторы постановили, что при выравнивании последовательностей проб и генов слишком много совпадений меньше, чем на 8 нуклеотидов, поэтому эти совпадения предлагается считать незначимыми. Чем меньше участок совпадения последовательностей, тем с меньшей вероятностью молекулы вступят во взаимодействие, поэтому авторы предполагают, что основная кросс-гибридизация будет происходить на участках от 10 до 16 совпадений. Участки совпадений длиннее 16 нуклеотидов встречаются крайне редко, так как пробы специально разрабатывались таким образом, чтобы вероятность кросс-гибридизации была как можно меньше.

В данной работе матрица A^{blast} была использована для оценки качества восстановления параметров модели, учитывающей кросс-гибридизацию. Для этого были

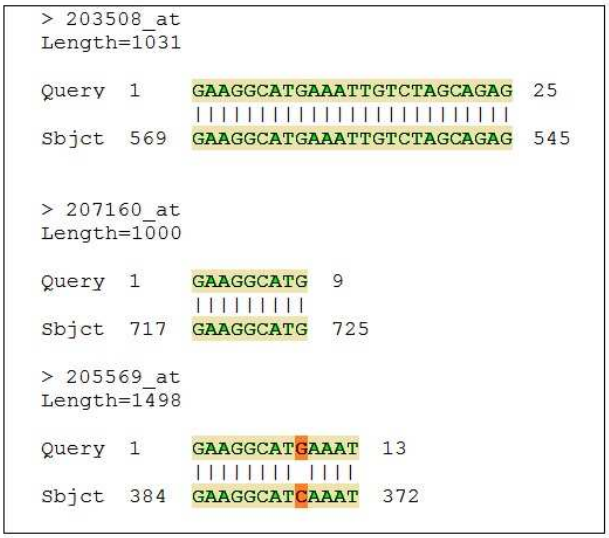


Рис. 7: Пример работы алгоритма BLASTN. Три типа ответа на запрос по одной пробе: полное совпадение (со специфическим геном), частичное совпадение (в середине), частичное совпадение с заменой одного нуклеотида (внизу)

введены два функционала качества, называемые функционалами ошибок I и II рода:

$$QA_I^{blast} = \frac{\sum_{ij}[A_{ij} = 0][A_{ij}^{blast} > 0]}{|A^{blast}|},$$

$$QA_{II}^{blast} = \frac{\sum_{ij}[A_{ij} > 0][A_{ij}^{blast} = 0]}{|A^{blast}|}.$$

Здесь $[\cdot]$ — индикатор истинности выражения, $|A^{blast}|$ — число элементов матрицы A. Отметим, что согласно сделанным предположениям о невозможности взамодействий некомплементарных молекул, функционал QA_{II}^{blast} можно использовать для оценки качества модели. Однако о функционале QA_I^{blast} конкретных выводов сделать нельзя, так как если проба частично комплементарна гену согласно алгоритму выравнивания, это ещё не значит, что она обязательно будет вступать с ним в реакцию, поэтому коэффициет взаимодействия может быть равным нулю.

3.7 Комбинированная модель

Предложенная выше модель кросс-гибридизации восстанавливает линейную зависимость между интенсивностью свечения проб и концентрацией генов. Wu et al. [29]

строили линейную модель кросс-гибридизации, потому что зависимость интенсивностей свечения проб от концентрации неспецифических генов хорошо описывается линейным законом. Ho Wu et al. рассматривали только неспецифические взаимодействия, которые не подчиняются закону насыщения. Многочисленные опыты показывают, что характер зависимости интенсивности свечения проб от концентрации специфических генов далеко не линейный. В работе Cambon et al. [8] попытались построить линейную модель данных, однако опыты показали, что смоделировать интенсивность свечения проб, учитывая только кросс-гибридизацию, с достаточной точностью не получается. При достаточно больших концентрациях молекул генов интенсивность свечения проб начинает подчиняться закону насыщения, описываемому в общем случае моделью Ленгмюра.

В данной работе предлагается скомбинировать модель кросс-гибридизации с моделью адсорбции Ленгмюра, подробно описанной выше. Зависимость интенсивности от концентрации генов с учетом насыщения будет выглядеть следующим образом:

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

α_i, β_i — параметры насыщения, γ_i — параметр неучтённой кросс-гибридизации, j_i — номер специфического гена для пробы i.

Нелинейная составляющая специфических взаимодействий немного видоизменена относительно базовой модели Ленгмюра. В этом виде модель позволяет легко описать линейную зависимость интенсивности от концентрации, если параметр β_i положить равным нулю. Кроме того, дополнительно вводится параметр неучтённых взаимодействий γ_i. Неучтённые взаимодействия могут возникнуть в том случае, если в образце присутствуют молекулы генов, не участвовавших при построении матрицы взаимодействий A, либо если по каким-то причинам произошла реакция пробы с некомплементарным геном или комплементарным менее, чем на 8 нуклеотидов.

Параметры модели $\alpha_i, \beta_i, \gamma_i, A = \{A_{ij}\}$ настраиваются одновременно в ходе минимизации функционала квадратичной невязки

$$Q = \sum_t \left(\sum_i (I_i^t - Y_i^t)^2 + \lambda \sum_{ij} |A_{ij}|\right) \to \min_{\alpha,\beta,\gamma,A}$$

где Y_i^t — наблюдаемая интенсивность i-й пробы на t-м чипе. λ — параметр регуляризации, может быть вектором (тогда каждый коэффициент будет иметь свой вес).

Коэффициенты матрицы A можно искать тем же способом, который был использован в модели кросс-гибридизации (с помощью L_1-регуляризации с параметром λ). Однако в предположениях о невозможности реакции в случае некомплементарности пробы и гена предлагается восстанавливать только те коэффициенты A_{ij}, для которых соответствующие коэффициенты $A_{ij}^{blast} > 0$. В этом случае можно обойтись без регуляризации.

Поиск исходных концентраций генов осуществляется в ходе минимизации функционала

$$Q = \sum_t \sum_i \left(I_i^t(C) - Y_i^t\right)^2 \to \min_C, \qquad C \geqslant 0. \tag{1}$$

В работе предлагается гибкая модель данных микрочипов ДНК, учитывающая линейные и нелинейные зависимости интенсивности от концентраций, специфические и неспецифические взаимодействия проб и генов. В разделе Эксперименты будут предоставлены результаты исследования параметров модели и сравнения её качества с существующими методами оценки экспрессии генов.

Часть III

Эксперименты

3.8 Описание исходных данных

Эксперимент «Латинский Квадрат»

Эксперимент «Латинский квадрат» был проведён на платформе Affymetrix специально для исследования и тестирования методов обработки данных микрочипов ДНК. В ходе эксперимента в образец с РНК клеточной линии HeLa (ATCC CCL-13) были добавлены РНК 42-х генов, изначально не содержащихся в данном образце. Эти 42 гена были выбраны на основании сборок генома человека GeneChip и Taqman.

Из этих 42-х генов 30 генов соответствуют комплементарным кДНК, выделенным из РНК клеточной линии лимфобластов. Ещё восемь генов имеют искусственную последовательность, до сих пор не обнаруженную в природе. Оставшиеся четыре

гена являются контрольными генами для микрочипов компании Affymetrix (входят в состав контрольного набора генов polyA).

Все гены были одинаковым образом выделены и подготовлены, а затем в известных концентрациях нанесены на микрочипы. Концентрация варьировалась в пределах значений $\{0, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$ пМ. Гены были сгруппированы по три в одной пробирке. Эксперимент состоял из 14-ти стадий, на каждой из которых на микрочип были нанесены смеси генов из разных пробирок в разных концентрациях. На каждой стадии одинаковые смеси были нанесены на три микрочипа (так называемые *технические репликаты*). Эти репликаты служат для контроля за воспроизводимостью эксперимента.

Таким образом, в нашем распоряжении оказались 42 микрочипа (14x3) с известными интенсивностями свечения проб и концентрациями 14-ти образцов (по три гена в каждом). Схема эксперимента изображена на рис.(8).

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

Рис. 8: Схема проведения эксперимента «Латинский квадрат». Матрица концентраций. Строки соответствуют экспериментам, столбцы — группе генов.

3.9 Анализ исходных данных

3.9.1 Согласованность репликатов

На каждой стадии эксперимента были изготовлены три технических репликата для контроля за воспроизводимостью экспериментов. Технические репликаты можно использовать также при анализе методов обработки данных микрочипов. В частности, репликаты можно использовать для оценки непротиворечивости модели, так

как заранее известно, что концентрации генов в нанесённых на репликаты смесях одинаковы.

В ходе анализа исходных данных была измерена согласованность интенсивностей свечения проб на репликатах. Для этого были использованы следующие оценки:

- Корреляция Пирсона

  Чтобы выделить пробы, интенсивность которых коррелирует между техническими репликатами и действительно соответствует изменению концентрации генов, мы решили воспользоваться методом Wu et al. [29]. Мы посчитали корреляцию интенсивностей проб на 1-м и 2-м и 1-м и 3-м репликатах для всего эксперимента (14 микрочипов). Если корреляция в обоих случаях не превышала 80%, авторы считали интенсивности проб шумовыми, так как интенсивности проб не воспроизводятся на репликатах. На Рис. 9 каждая точка соответствует пробе, координаты точки - корреляциям интенсивностей для 1-го и 2-го и 1-го и 3-го репликатов соответственно. Интенсивности точек, лежащих внутри этого круга, авторы предлагают не учитывать при анализе модели. В правом верхнем углу располагаются пробы, соответствующие генам, добавленным в эксперименте «Латинский квадрат», — у них наиболее высокая корреляция между репликатами.

- Модель линейной регрессии

  В этой модели интенсивность проб на одном чипе линейно выражается через интенсивность проб на другом чипе: $I_i^{t_1} = \alpha I_i^{t_2} + \beta$, коэффициенты α и β были найдены в ходе минимизации квадратичной невязки

  $$RSS = \sum_i \left(I_i^{t_1} - (\alpha I_i^{t_2} + \beta)\right)^2 \to \min_{\alpha,\beta}.$$

  Результаты сравнения интенсивности свечения микрочипов представлены на Рис. 10-11. На Рис. 10 изображены значения функционала RSS для всех микрочипов из эксперимента. Можно заметить преобладание низких значений функционала RSS на трёх диагоналях (тёмно-синяя полоса на рисунке). Это соответствует согласованности технических репликатов. На Рис. 11 изображены значения функционала RSS для 2-го и 7-го экспериментов (по три технических репликата в каждом). Проведённое исследование подтверждает согласо-

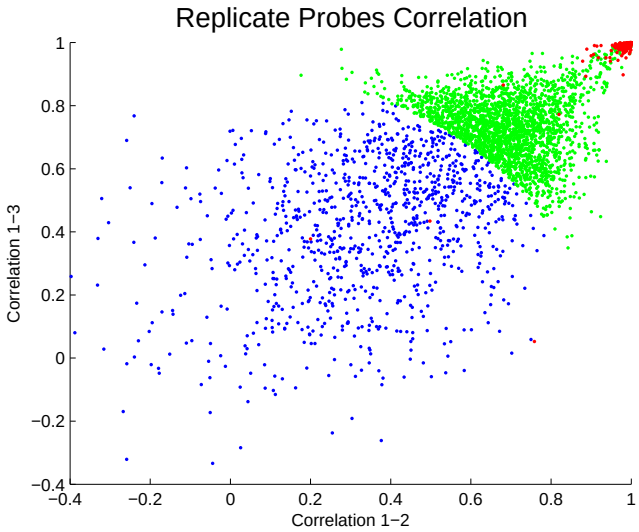


Рис. 9: Корреляция интенсивностей проб между репликатами. По оси X отложено значение корреляции между 1-м и 2-м, по оси Y — между 1-м и 3-м репликатами. Зелёным цветом отмечены пробы, лежащие вне круга радиуса 0.9 (высоко коррелированные по обоим репликатам). Красным —пробы, соответствующие генам известной концентрации из эксперимента «Латинский квадрат».

ванность данных на репликатах и возможность использования последних для оценки устойчивости модели.

3.9.2 Влияние комплементарности на гибридизацию

В данном разделе мы попытались наглядно показать, действительно ли комплементарность пробы неспецифическому гену может быть причиной кросс-гибридизации и влиять на интенсивность свечения пробы при изменении концентрации неспецифического гена. На Рис. 12 изображены значения интенсивностей проб на двух чипах, на одном из которых (по оси Y) один ген отсутствует, а на другом (по оси X) содержится в концентрации 4пМ. Цвет точек соответствует коэффициенту комплементарности пробы к данному гену. Как можно заметить, интенсивность проб, частично комплементарных данному гену, увеличивается при добавлении гена в образец. Это ободряет нас в использовании матрицы комплементарности при стремлении учесть эффект кросс-гибридизации в модели данных микрочипов ДНК.

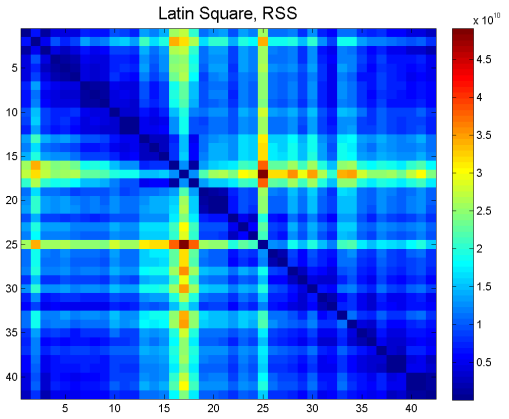


Рис. 10: Величина коэффициента RSS в сравнении микрочипов из эксперимента «Латинский квадрат»

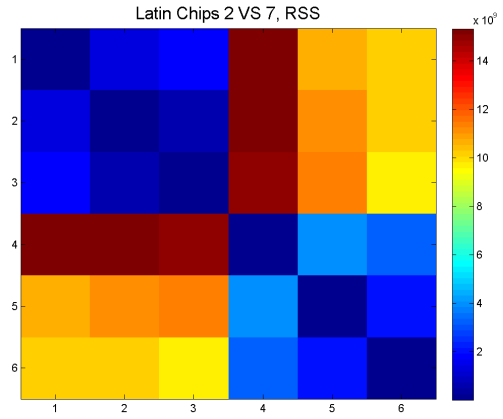


Рис. 11: Величина коэффициента RSS в сравнении микрочипов (2,7) из эксперимента «Латинский квадрат»

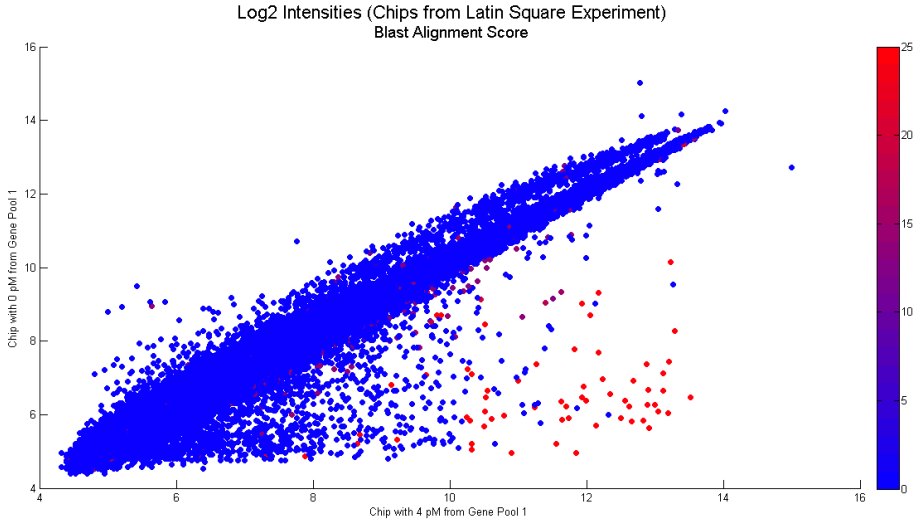


Рис. 12: Интенсивность свечения проб на микрочипе при отсутствии гена (по оси Y) и при концентрации 4пМ (по оси X). Цвет точки соответствует величине коэффициента копмлементарнности пробы гену (BLASTN)

3.10 Восстановление матрицы взаимодействий

3.10.1 Функционалы качества модели

Для оценки качества модели и сравнения моделей между собой были предложены несколько функционалов качества.

Оценка устойчивости модели Для оценки устойчивости модели была использована информация о технических репликатах. На микрочипы-репликаты были нанесены одинаковые образцы, поэтому восстанавливаемые концентрации и оцениваемые коэффициенты матрицы A должны быть как можно более близки. Для сравнения модельных матриц взаимодействий использовался следующий функционал качества:

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

где $t_1 < t_2$ - номер репликата, $t_1, t_2 \in \{1, 2, 3\}$. Для сравнения восстановленных матриц концентраций был предложен функционал качества

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

Результаты экспериментов по восстановлению матриц взаимодействий и концентраций изображены на Рис. 13-20.

Оценка качества модели Для оценки качества модели восстановленные концентрации сравнивались с изначально известными концентрациями генов эксперимента «Латинский квадрат». Для этого был введён следующий функционал качества:

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

Здесь $t \in \{1, 2, 3\}$ — номер технического репликата, C — матрица концентраций, восстановленная для t-го репликата эксперимента, $\tilde{C}$ — матрица известных концентраций.

Для определения качества восстановления коэффициентов взаимодействия проб и генов мы воспользовались матрицей A^{blast}, построенной с помощью алгоритма выравнивания символьных последовательностей BLASTN. Функционалы качества выглядят следующим образом:

$$QA_I^{blast} = \frac{1}{3}\sum_{t}\frac{\sum_{ij}[A_{ij}^{t} = 0][A_{ij}^{blast} > 0]}{|A^{blast}|},$$

$$QA_{II}^{blast} = \frac{1}{3}\sum_{t}\frac{\sum_{ij}[A_{ij}^{t} > 0][A_{ij}^{blast} = 0]}{|A^{blast}|}.$$

Здесь A^t — матрица коэффициентов взаимодействий проб и генов, построенная для t-го репликата эксперимента. Как уже отмечалось ранее, функционал QA_{II}^{blast} меряет ошибки построенной модели, так как в силу предположения о невозможности некомплементарных взаимодействий наша модель допускает ошибку, если восстановленный коэффициент взаимодействия некомплементарных пробы и гена отличен он нуля. Функционал QA_{I}^{blast}, однако, нельзя назвать функционалом ошибок, так как частично комплементарные пробы и ген могут как вступать, так и не вступать во взаимодействия, и восстановленный коэффициент может оказаться равным нулю. Этот функционал показывает близость восстановленной матрицы к матрице A^{blast} и, скорее, может считаться оценкой коэффициентов матрицы A^{blast}.

3.10.2 Регуляризация

Ввиду того, что число восстанавливаемых параметров велико (размерность матрицы взаимодействий 996×14), может возникнуть нежелательный эффект *переобучения*, когда модель слишком хорошо настраивается на входные данные. Для того, чтобы избежать переобучения и контролировать коэффициенты матрицы взаимодействий A, было предложено использовать регуляризацию. Предполагалось, что регуляризация позволит избежать возникновения ненулевых коэффициентов на тех позициях, где взаимодействие пробы и гена невозможно в природе. В этой секции представлены результаты исследования влияния регуляризации на качество модели.

L_1-регуляризация с одним параметром λ Вначале была рассмотрена L_1-регуляризация с одним параметром λ для всех коэффициентов матрицы A. При этом были рассмотрены два способа восстановления матрицы A: построчно (задача минимизации для каждой пробы отдельно) и целиком (общая задача восстановления A). Минимизируемый функционал в первом случае имеет вид:

$$Q_i = \sum_t (I_i^t - Y_i^t)^2 + \lambda_i \sum_j |A_{ij}| \to \min_A$$

Здесь I_i^t — восстановленная интенсивность i-й пробы на t-м чипе, Y_i^t — наблюдаемая интенсивность, A_{ij} — коэффициент взаимодействия i-й пробы с j-м геном. Параметр регуляризации λ_i задаётся отдельно для каждой подзадачи для i-й пробы.

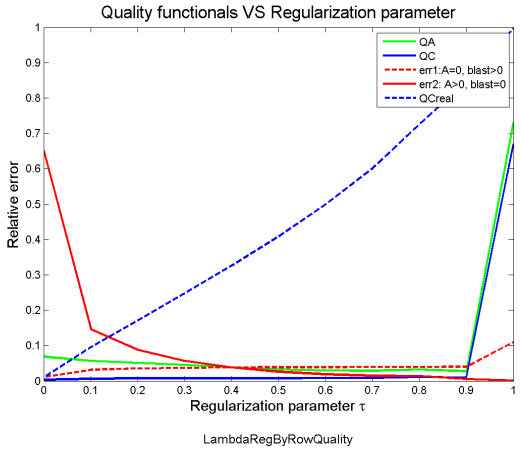


Рис. 13: Функционалы качества в зависимости от величины параметра регуляризации λ для построчного восстановления A

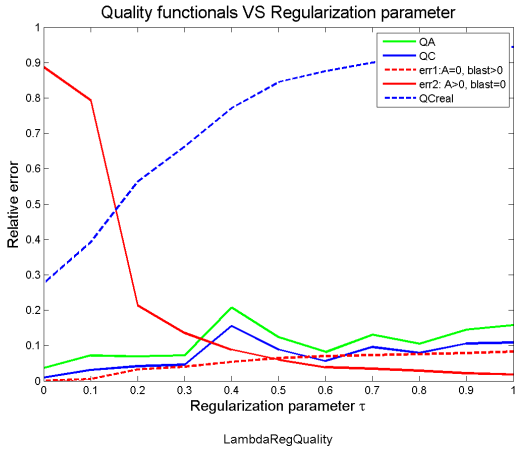


Рис. 14: Функционалы качества в зависимости от величины параметра регуляризации λ, восстановление A целиком

Для задачи восстановления матрицы A целиком функционал будет иметь вид:

$$Q = \sum_t \left(\sum_i (I_i^t - Y_i^t)^2 \right) + \lambda \sum_{ij} |A_{ij}| \to \min_A$$

В этом случае суммирование ведётся по всем пробам i на всех микрочипах t, регуляризация затрагивает все коэффициенты матрицы A (для всех проб i и всех генов j).

На Рис. 13-14 можно заметить, что модель достаточно устойчива: функционалы QA и QC (на графике обозначены сплошными зелёной и синей линиями) изменяются в пределах 10% для построчного восстановления A и 20% для восстановления A целиком. Однако увеличение параметра регуляризации ведёт к уменьшению устойчивости. Регуляризация помогает уменьшить функционал QA_{II}^{blast} (сплошная красная линия), однако резко увеличивается разница с исходными концентрациями генов (QC^{real}, пунктирная синяя линия).

Специализированная L_1-регуляризация с параметром λ В том случае, когда регуляризация затрагивает все коэффициенты матрицы A, неизбежно уменьшаются не только коэффициенты неспецифических взаимодействий, но и коэффициенты специфических взаимодействий. Нашей задачей было уменьшить лишь коэффициенты неспецифических взаимодействий. Для того, чтобы не «наказывать» коэффициенты

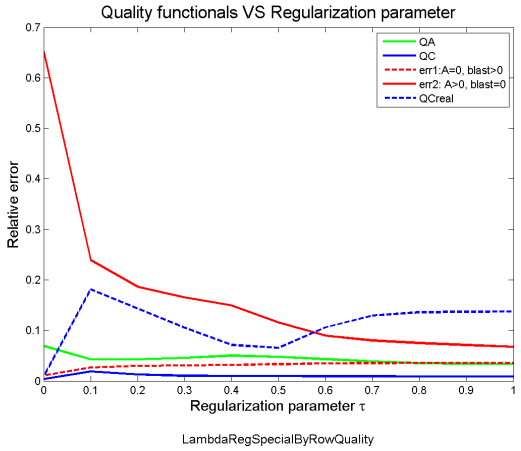


Рис. 15: Функционалы качества в зависимости от величины параметра специализированной регуляризации λ для построчного восстановления A

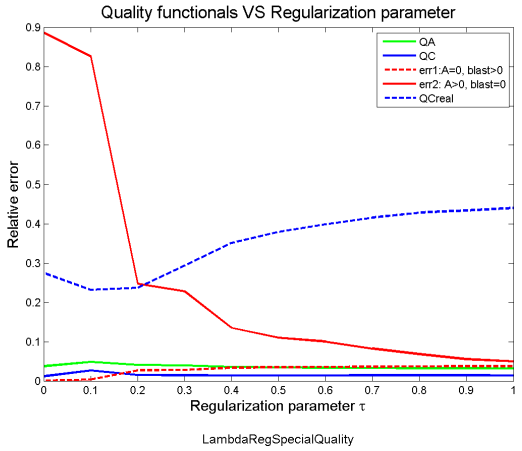


Рис. 16: Функционалы качества в зависимости от величины параметра специализированной регуляризации λ, восстановление A целиком

специфических взаимодействий (назовём их *ведущими*), была предложена специализированная L_1-регуляризация. В этом случае функционалы имеют вид:

$$Q_i = \sum_t (I_i^t - Y_i^t)^2 + \lambda_i \sum_{j, j \neq j_0^i} |A_{ij}| \to \min_A$$

$$Q = \sum_t \left(\sum_i (I_i^t - Y_i^t)^2 \right) + \lambda \sum_{ij, j \neq j_0^i} |A_{ij}| \to \min_A$$

Здесь суммирование в слагаемом регуляризации идёт по всем коэффициентам A_{ij} за исключением ведущих коэффициентов $A_{ij_0^i}$. Специализированная регуляризация была осуществлена следующим образом: сначала без регуляризации были найдены ведущие коэффициенты $A_{ij_0^i}$, после чего размерность задачи была понижена и задача была решена только для неспецифических коэффициентов.

На Рис. 15-16 очевидно преимущество специализированной регуляризации над обычной. Функционал QC^{real} разницы восстановленных и истинных концентраций изменяется в пределах 20% для построчного восстановления A и 40% для восстановления целиком. Однако качество восстановленных концентраций оставляет желать лучшего, кроме того, значение функционала QA_{II}^{blast} велико — $20 - 30\%$.

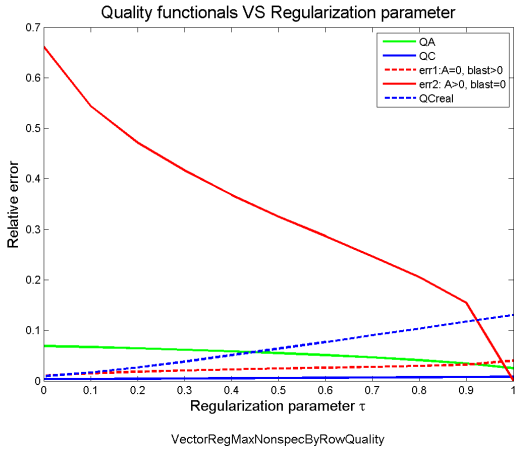


Рис. 17: Функционалы качества в зависимости от величины параметра векторной регуляризации λ для построчного восстановления A

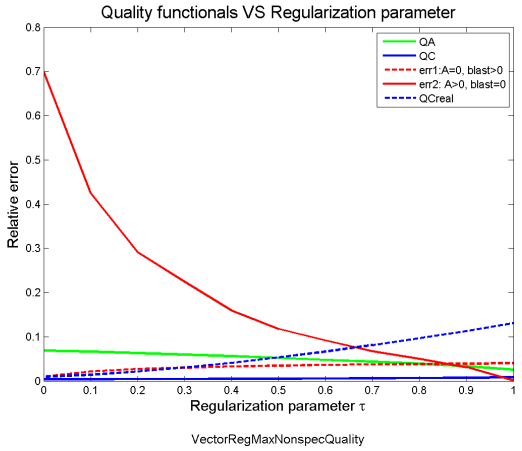


Рис. 18: Функционалы качества в зависимости от величины параметра векторной регуляризации λ, восстановление A целиком

L_1-регуляризация с вектором параметров $\vec{\lambda}$ Главным недостатком специализированной L_1-регуляризации, предложенной выше, является то, что ведущие коэффициенты не изменяются. При решении задачи меньшей размерности уменьшаются лишь неспецифические коэффициенты, и от этого страдает точность приближения. Поэтому далее была рассмотрена L_1-регуляризация с вектором параметров $\vec{\lambda}$. В этом случае минимизируемые функциналы имеют вид

$$Q_i = \sum_t (I_i^t - Y_i^t)^2 + \sum_j \lambda_{ij}|A_{ij}| \to \min_A$$

$$Q = \sum_t \left(\sum_i (I_i^t - Y_i^t)^2 \right) + \sum_{ij} \lambda_{ij}|A_{ij}| \to \min_A, \qquad \lambda_{ij_0^i} = 0$$

Преимущество регуляризации с вектором параметров в том, что в то время как неспецифические коэффициенты должны уменьшаться, ведущие могут возрастать.

На Рис. 17-18 изображены функционалы качества модели в зависимости от параметра векторной регуляризации. Векторная регуляризация имеет существенное преимущество перед скалярной, потому что параметры модели настраиваются более гибко. Разница восстановленных концентраций и настоящих относительно невелика: в пределах 10% для построчного восстановления A и 20% для восстановления матрицы A целиком. Хотя при увеличении параметра регуляризации уменьшается ошиб-

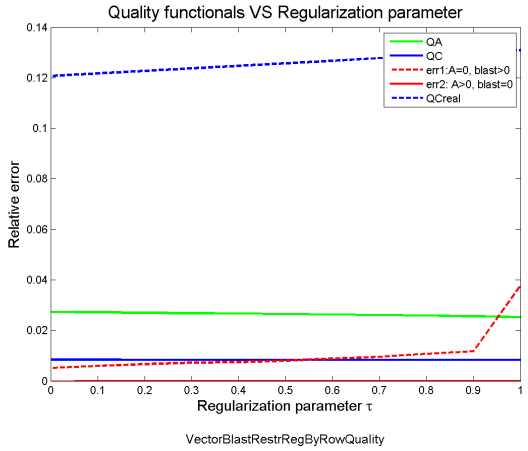


Рис. 19: Функционалы качества в зависимости от величины параметра векторной регуляризации λ для построчного восстановления A, ограничения Blast

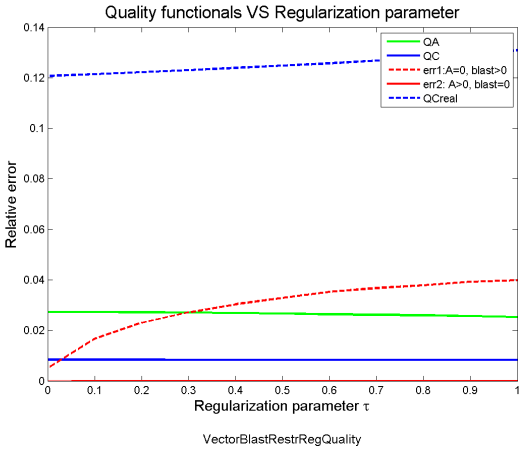


Рис. 20: Функционалы качества в зависимости от величины параметра векторной регуляризации λ, восстановление A целиком, ограничения Blast

ка второго рода (функционал QA_{II}^{blast}), то есть уменьшается количество ненулевых коэффициентов, которые должны быть нулевыми из биологических соображений, существенные преимущества применения регуляризации не выявлены.

Для сравнения на Рис. 19-20 изображены функционалы качества модели в зависимости от параметра векторной регуляризации в том случае, когда в матрице A восстанавливались коэффициенты взаимодействий только хотя бы частично комплементарных проб и генов (согласно алгоритму BLASTN). Качество восстановления концентраций сравнительно хорошее — ошибка около 12%. Регуляризация в этом случае оказывается уже не нужна, так как функционал QA_{II}^{blast} по построению модели равен нулю, а значения остальных функционалов с увеличением параметра регуляризации только возрастают.

Исследования влияния регуляризации на качество восстановленной модели показали, что регуляризация позволяет сократить переобучение, то есть восстановление заведомо нулевых коэффициентов взаимодействия проб и генов. Из нескольких модификаций регуляризации лучше всего показала себя векторная, при которой все коэффициенты матрицы A получали свой вес. Таким образом можно регулировать

увеличение специфических коэффициентов и уменьшение других. Однако схожее качество восстановления концентраций можно получить, если изначально наложить на матрицу A ограничения, получаемые с помощью матрицы комплементарности алгоритма BLASTN. Поэтому для восстановления концентраций в комбинированной модели, учитывающей закон насыщения, в дальнейшем использовалась матрица A, полученная с помощью ограничений матрицы A^{blast}.

3.10.3 Сравнение восстановления построчно и целиком

В экспериментах по исследованию влияния регуляризации на качество модели использовалось как построчное восстановление матрицы A, так и восстановление A целиком. Сделать однозначный вывод о том, какой метод предпочтительней, сложно. Основное преимущество построчного восстановления матрицы A — резкое сокращение размерности задачи. В этом случае ищутся коэффициенты взаимодействия одной пробы со всеми генами, которые могут быть в образце. Построчное восстановление возможно из соображений независимости гибридизации проб: мы предполагаем, что молекул гена в образце настолько много, что они могут связаться со всеми пробами на чипе. На практике это требование выполняется, так как известно, что на этапе промывки чипа удаляется около 90% образца с негибридизовавшимися молекулами генов. Построчное восстановление может сыграть существенную роль при увеличении размерности задачи.

Однако в случае, когда размерность задачи позволяет восстанавливать матрицу A целиком, мы получаем существенный выигрыш во времени работы программы. Вместо решения нескольких сотен или тысяч задач мы решаем всего одну. Кроме того, при восстановлении матрицы A целиком регуляризация применяется ко всем коэффициентам одновременно, а не только к коэффициентам взаимодействия одной пробы. Поэтому в этом случае регуляризация более точно решает поставленную задачу сокращения неспецифических коэффициентов и поощрения специфических и коэффициентов кросс-гибридизации. На Рис. 21 изображены упорядоченные коэффициенты матрицы A, восстановленые построчно (зеленые) и одновременно (фиолетовые). Заметна тенденция увеличения маленьких коэффициентов для построчного восстановления и увеличения больших для восстановления целиком. Это объясняется

тем, что при восстановлении матрицы A построчно при решении каждой подзадачи рассматриваются коэффициенты только одной пробы, и их значения распределяются более равномерно в строке матрицы A, чем при восстановлении A целиком.

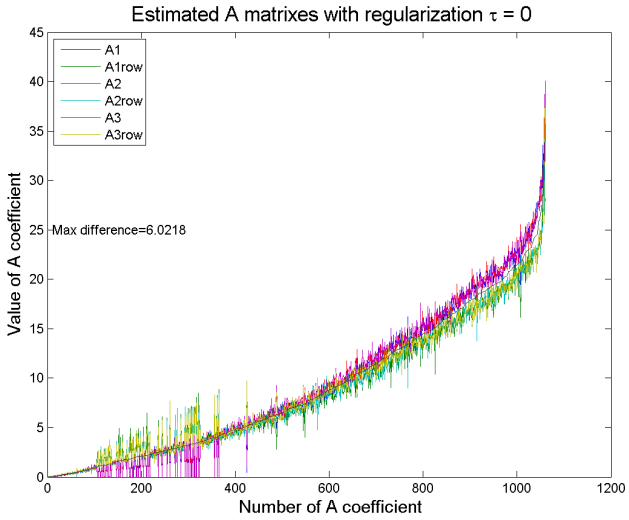


Рис. 21: Упорядоченные коэффициенты матрицы , восстановленной построчно (зелено-желтые) и целиком (сине-фиолетовые). Заметна тенденция к увеличению маленьких коэффициентов и уменьшения больших для построчного восстановления A. (Без регуляризации.) По оси X отложен порядковый номер коэффициента.

В дальнейшем выбор метода восстановления A (построчно или целиком) предлагается на усмотрение исследователя в зависимости от размерности задачи и ограничения временных затрат.

3.11 Сравнение репликатов

В эксперименте «Латинский квадрат» для каждой из 14-ти стадий эксперимента были поставлены три микрочипа — технических репликата. Таким образом, можно 3^{14} способами выбрать микрочипы для исследования одного эксперимента. Мы решили удостовериться в том, что выбор репликатов не влияет на восстановление матрицы A.

Были сгенерированы 100 случайных разбиений выборки на репликаты для трёх экспериментов. Таким образом, мы получили 300 репликатов эксперимента из 14-ти стадий. Для каждого эксперимента была восстановлена матрица A методом ми-

нимизации квадратичной невязки с векторной L_1-регуляризацией для разных значений параметра регуляризации $\vec{\lambda}$. Матрица A была восстановлена также по полной выборке из 42-х микрочипов.

Относительная вариация коэффициентов матрицы A для каждого значения параметра регуляризации $\vec{\lambda}$ была посчитана по формуле

$$RelVar(A) = \frac{\sum_{ij}(\max A_{ij} - \min A_{ij})}{\sum_{ij} \tilde{A}_{ij}}$$

Здесь $\tilde{A}$ — матрица, восстановленная по всей выборке из 42-х чипов, максимум и минимум берутся по всем восстановленным матрицам. График зависимости величины относительной вариации коэффициентов матрицы A от величины параметра регуляризации λ изображен на Рис. 22. Вариация коэффициентов уменьшается с увеличением регуляризации, что вызвано накладыванием более строгих ограничений на неспецифические коэффициенты. На Рис. 23 изображены величины коэффициентов матрицы A для случайных 30 репликатов при восстановлении матрицы A без регуляризации. В среднем вариация составляет 10%, из чего можно сделать вывод, что выбор репликатов не оказывает существенного влияния на восстановленую матрицу A.

Кроме того, было проведено сравнение коэффициентов матрицы A, построенной по 1-му, 2-м и 3-м репликатам одновременно. График относительной вариации коэффициентов этих матриц в зависимости от величины параметра регуляризации изображен на Рис. 24.

Изменение вариации в этом случае совсем незначительно, в пределах 1%. На Рис. 25 изображены упорядоченные коэффициенты матрицы A построенной по 1-му, 2-м и 3-м репликатам одновременно. Этот график подтверждает то, что вариация коэффициентов очень мала.

Проведённое исследование показывает, что предлагаемая модель достаточно устойчива и её коэффициенты не зависят от выбранного репликата эксперимента.

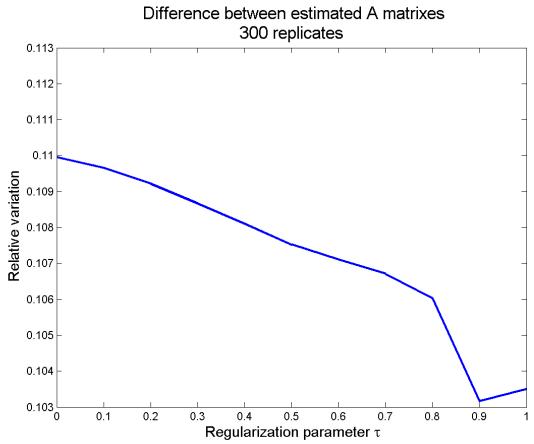


Рис. 22: Относительная вариация коэффициентов матрицы A, построенной по случайным 300-м репликатам в зависимости от величины параметра регуляризации.

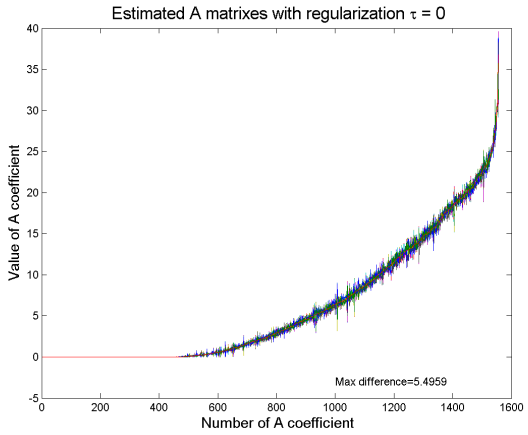


Рис. 23: Упорядоченные коэффициенты матрицы , полученной при восстановлении по случайным 30-ти репликатам без регуляризации. По оси X отложен порядковый номер коэффициента.

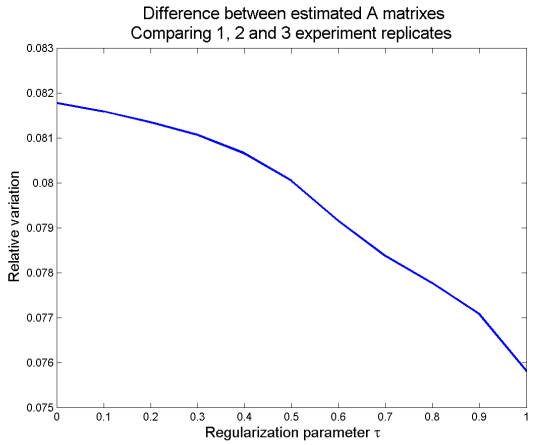


Рис. 24: Относительная вариация коэффициентов матрицы A, построенной по одному, двум и трём репликатам одновременно в зависимости от величины параметра регуляризации.

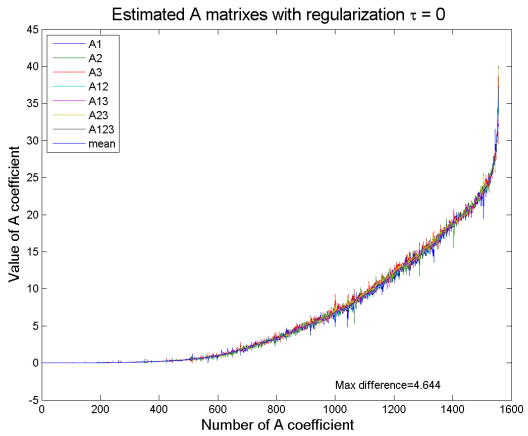


Рис. 25: Упорядоченные коэффициенты матрицы , полученной при восстановлении по одному, двум и трём репликатам без регуляризации. По оси X отложен порядковый номер коэффициента.

3.12 Восстановление концентраций генов

Основной задачей анализа данных микрочипов ДНК является восстановление исходных концентраций генов. Для этого была предложена комбинированная модель, учитывающая эффекты кросс-гибридизации и насыщения. При восстановлении исходных концентраций использовалась матрица взаимодействий A, полученная с ограничениями, накладываемыми матрицей A^{blast}. Графики восстановленных концентраций для нескольких групп генов эксперимента «Латинский квадрат» изображены на Рис. 26.

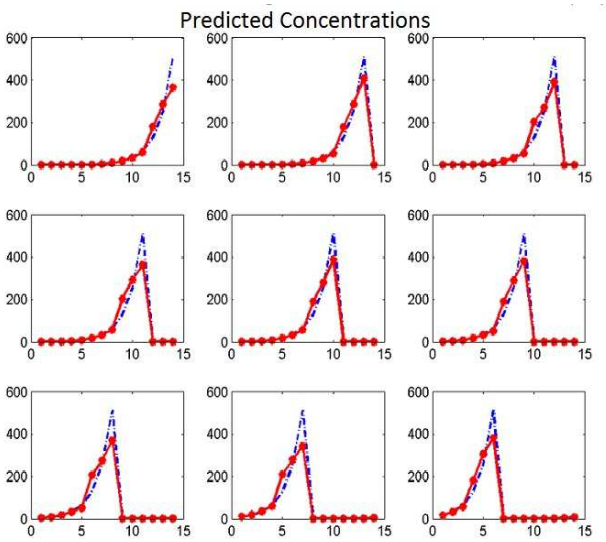


Рис. 26: Восстановленные концентрации генов некоторых генов эксперимента «Латинский квадрат» комбинированной моделью. Синим цветом обозначены истинные концентрации, красным — восстановленные.

Можно заметить, что модель недооценивает высокие концентрации генов. Чуть позже будет приведено сравнение результатов работы предложенной модели с другими методами.

Для оценки качества модели и проверки на переобучение был проведён следующий эксперимент. По одному из репликатов была восстановлена матрица взаимодействий A. Далее концентрации генов были восстановлены с использованием матрицы A для этого же репликата (то есть, на обучающей выборке) и для других репликатов (на так называемой контрольной выборке). Обучающий репликат эксперимента был выбран трижды и построены графики распределения восстановленных

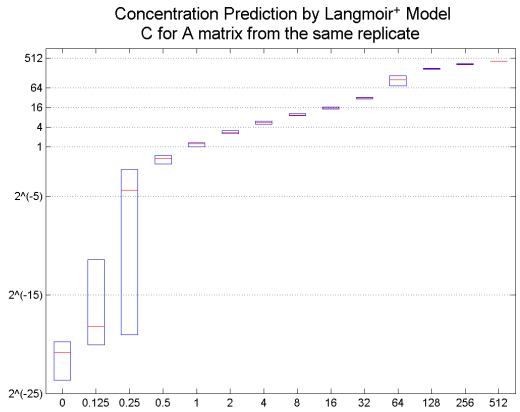


Рис. 27: Разброс восстановленных концентраций на обучающей выборке (матрица A была построена по тому же репликату)

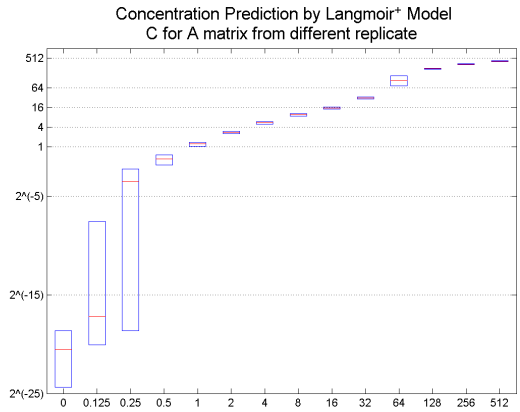


Рис. 28: Разброс восстановленных концентраций на контрольной выборке (матрица A была построена по другому репликату)

концентраций (Рис. 27-28). На графиках видно, что модель достаточно устойчива и разброс восстановленных концентраций мал (за ислючением небольших концентраций).

Значения небольших концентраций имеют большой разброс в том числе потому, что концентрации представлены в логарифмической шкале, и очень маленькие близкие к нулю значения могут иметь разброс $10^{-5} - 10^{-15}$. В целом можно сделать вывод об устойчивости модели и об отсутствии переобучения.

3.12.1 Сравнение с существующими методами

Было проведено сравнение предложенной комбинированной модели с распространёнными существующими методами MAS 5.0, RMA, DFCM и FARMS. Более подробное описание этих методов см. в таблице 3.1. Основное отличие предложенной модели от этих методов в том, что она учитывает биологическую вариацию и ориентирована на восстановление исходных концентраций. Рассмотренные методы больше направлены на определение изменения уровня экспрессии.

На Рис. 30 изображены восстановленные изменения уровня экспрессии относительно единицы измерения. Чёрная линия обозначает истинные изменения уровня

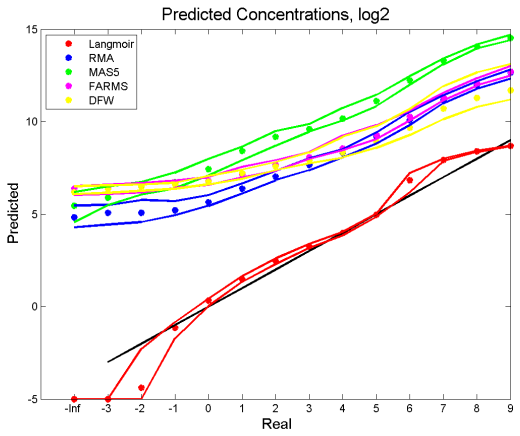


Рис. 29: Восстановленные концентрации ($\log_2 C$). Комбинированная модель обозначена красным, истинные значения - чёрным

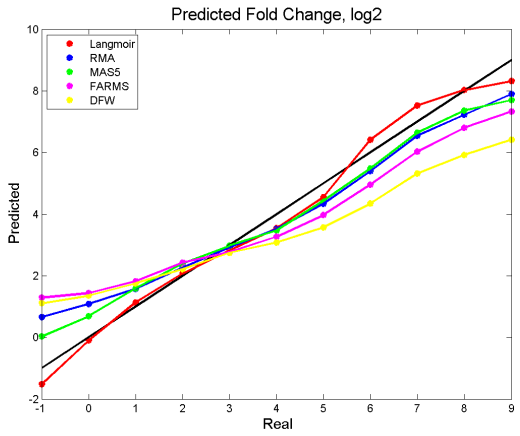


Рис. 30: Восстановленные изменения концентрации (Fold Change). Комбинированная модель обозначена красным, истинные значения - чёрным

экспрессии. Как можно заметить, комбинированная модель предсказывает изменение уровня экспрессии не хуже, чем распространённые методы. Более того, в отличие от остальных методов, предложенная модель не имеет тенденции завышать оценку для маленьких концентраций и занижать её для больших.

На Рис. 29 изображены восстановленные логарифмированные концентрации, чёрная линия — истинные значения. Распространённые методы, не ориентированные на точное восстановление концентраций, очень сильно завышают значения концентраций. Предложенная модель, наоборот, достаточно хорошо приближает исходные концентрации.

Таким образом, можно сделать вывод, что предложенная комбинированная модель, учитывающая кросс-гибридизаию и эффект насыщения, более предпочтительна как для определения исходных концентраций, так и для оценки изменения уровня экспрессии.

Часть IV

Заключение

В настоящее время стремительно развиваются технологии в области биологии и биоинформатики, и перед исследователями встают новые задачи, требующие детального изучения. В данной работе была исследована актуальная проблема анализа данных микрочипов ДНК и рассмотрены различные подходы к её решению.

Основные результаты работы:

- Построены модели данных микрочипов ДНК, учитывающие специфические и неспецифические взаимодействия молекул на поверхности чипа.
- Исследовано влияние регуляризации на качество линейной модели, учитывающей неспецифические взаимодействия. Относительные погрешности построенной модели оказались достаточно малы — в пределах 10%. Адекватность модели с точки зрения биологических соображений остаётся предметом исследования.
- Построена комбинированная нелинейная модель, позволяющая учесть разный характер зависимости в данных, а также биологические и технические вариации. Произведено сравнение модели с существующими методами на экспериментальных данных.
- Предложенная модель оценивает изменение уровня экспрессии генов, как минимум, не хуже существующих широко распространённых методов анализа данных микрочипов ДНК.
- При восстановлении исходных концентраций генов предложенная комбинированная модель даёт более точные результаты, чем популярные существующие методы.

Предложенная модель направлена на решение задачи исследования экспрессии генов в образцах на разных стадиях эксперимента. В ней были учтены эффекты кросс-гибридизации и насыщения, что позволяет приближать исходные концентрации генов и сравнивать экспрессию разных генов или даже результаты, полученные в ходе других экспериментов.

Предполагается проверять результаты работы предложенной модели на данных, полученных с помощью других методов оценки экспрессии (количественной ПЦР и секвенирования РНК). В будущем планируется получение новых данных, располагая которыми возможно будет оценивать, настраивать и модифицировать предлагаемые модели с целью повышения качества анализа данных.

Автор работы выражает благодарность сотрудникам лаборатории Молекулярной Физиологии при ВНИИ Физической Культуры и Спорта, возглавляемой к.б.н. Д.А. Сахаровым. Отдельную благодарность автор выражает Евгению Рябенко, предоставлявшему данные и делившемуся опытом в сфере биоинформатики и анализа микрочипов ДНК, и научному руководителю К.В. Воронцову.

Список литературы

[1] *Abdueva D., Skvortsov D., Tavare S.* Non-linear analysis of GeneChip arrays // *Nucleic acids research.* — 2006.— Vol. 34, no. 15. — Pp. e105.

[2] *Affymetrix Inc.* Statistical Algorithms Description Document (http://www.affymetrix.com/support/technical/whitepapers.affx)

[3] *Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J.* Basic local alignment search tool // *Journal of Molecular Biology.* — 1990. — Vol. 215, no. 3. — Pp. 403-410.

[4] *Bolstad B.M., Irizarry R.A., Astrand M., Speed T.P.* A comparison of normalization methods for high density oligonucleotide array data based on variance and bias // *Bioinformatics.* — 2003. — Vol. 19, no. 2. — Pp. 185-193.

[5] *Burden C.J., Binder H.* Physico-chemical modelling of target depletion during hybridization on oligonulceotide microarrays // *Physical Biology.* — 2009. — Vol. 7, no. 1. — Pp. 016004.

[6] *Burden C.J., Pittelkow Y.E., Wilson S.R.* Statistical Analysis of Adsorption Models for Oligonucleotide Microarrays // *Statistical Applications in Genetics and Molecular Biology.* — 2004. — Vol. 3, no. 1. — Article 35.

[7] *Burden C.J., Pittelkow Y.E., Wilson S.R.* Adsorption models of hybridisation and post-hybridisation behaviour on oligonucleotide microarrays // *Journal of Physics: Condensed Matter.* — 2006. — Vol. 23, no. 18. — Pp. 1-26.

[8] *Cambon A.C., Khalyfa A., Cooper N.G., Thompson C.M.* Analysis of probe level patterns in Affymetrix microarray data // *BMC bioinformatics.* — 2007. — Vol. 8, no. 1. — Pp. 146-156.

[9] *Chen Z., McGee M., Liu Q.* A Distribution-Free Convolution Model for background correction of oligonucleotide microarray data // *BMC genomics.* — 2009. — Vol. 10. — Pp. S19.

[10] *Eklund A.C., Friis P., Wernersson R., Szallasi Z.* Optimization of the BLASTN substitution matrix for prediction of non-specific DNA microarray hybridization // *Nucleic acids research.* — 2010. — Vol. 38, no. 4. — Pp. e27.

[11] *Furusawa C., Ono N., Suzuki S., Agata T., Shimizu H., Yomo T.* Model-based analysis of non-specific binding for background correction of high-density oligonucleotide microarrays // *Bioinformatics.* — 2009. — Vol. 25, no. 1. — Pp. 36-41.

[12] *Halperin A., Buhot A., Zhulina E.B.* Sensitivity, specificity, and the hybridization isotherms of DNA chips // *Biophysical journal.* — 2004. — Vol. 86, no. 2. — Pp. 718-730.

[13] *Heim T., Tranchevent L.C., Carlon E., Barkema G.T.* Physical-Chemistry-Based Analysis of Affymetrix Microarray Data // *Journal of Physical Chemistry B.* — 2006. — Vol. 110, no. 45. — Pp. 22786–22795.

[14] *Hekstra D., Taussig A.R., Magnasco M., Naef F.* Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays // *Nucleic Acids Research.* — 2003. — Vol. 31, no. 7. — Pp. 1962-1968.

[15] *Held G.A., Grinstein G., Tu Y.* Modeling of DNA microarray data by using physical properties of hybridization // *Proceedings of the National Academy of Sciences of the United States of America.* — 2003. — Vol. 100, no. 13. — Pp. 7575-7580.

[16] *Held G.A., Grinstein G., Tu Y.* Relationship between gene expression and observed intensities in DNA microarrays — a modeling study // *Nucleic Acids Research.* — 2006. — Vol. 34, no. 9. — Pp. e70.

[17] *Hochreiter S., Clevert D.A., Obermayer K.* A new summarization method for Affymetrix probe level data // *Bioinformatics.* — 2006. — Vol. 22, no. 8. — Pp. 943-949.

[18] *Irizarry R.A., Hobbs B., Collin F., et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data // *Biostatistics.* — 2003. — Vol. 4, no. 2. — Pp. 249-264.

[19] *Koltai H., Weingarten-Baror C.* Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction // *Nucleic Acids Research*. — 2008. — Vol. 36, no. 7. — Pp. 2395-2405.

[20] *Langdon W.B., Upton G.J., Harrison A.P.* Probes containing runs of guanines provide insights into the biophysics and bioinformatics of Affymetrix GeneChips // *Briefings in bioinformatics*. — 2009. — Vol. 10, no. 3. — Pp. 259-277.

[21] *Li C., Wong W.H.* Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection // *Proceedings of the National Academy of Sciences of the United States of America*. — 2001. — Vol. 98, no. 1. — Pp. 31-36.

[22] *Luebke K.J., Balog R.P., Garner H.R.* Prioritized selection of oligodeoxyribonucleotide probes for efficient hybridization to RNA transcripts // *Nucleic Acids Research*. — 2003. — Vol. 31, no. 2. — Pp. 750-758.

[23] *Marioni J.C., Mason C.E., Mane S.M., Stephens M., Gilad Y.* RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays // *Genome Research*. — 2008. — Vol. 18, no. 9. — Pp. 1509-1517.

[24] *Mulders G.C., Barkema G.T., Carlon E.* Inverse Langmuir method for oligonucleotide microarray analysis // *BMC Bioinformatics*. — 2009. — Vol. 10. — Pp. 64-72.

[25] *Naef F., Magnasco M.* Solving the riddle of the brigH mismatches: labeling and effective binding in oligonucleotide arrays // *Physical Review*. — 2003. — Vol. 68, no. 1. — Pp. 1-4.

[26] *Nguyen K.* Extended investigations in the physics of oligonucleotide microarrays. *The ANU Undergraduate Research Journal Volume One 2009*. — 2009. — no. 1. — Pp. 29-38.

[27] *Ono N., Suzuki S., et al.* An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays // *Bioinformatics*. — 2008. — Vol. 24, no. 10. — Pp. 1278-1285.

[28] *Wang Z., Gerstein M., Snyder M.* RNA-Seq: a revolutionary tool for transcriptomics // *Nature Reviews Genetics*. — 2009. — Vol. 10, no. 1. — Pp. 57-63.

[29] *Wu C., Carta R., Zhang L.* Sequence dependence of cross-hybridization on short oligo microarrays // *Nucleic acids research*. — 2005. — Vol. 33, no. 9. — Pp. e84.

[30] *Zhang L., Miles M.P., Aldape K.D.* A model of molecular interactions on short oligonucleotide microarrays // *Nature biotechnology*. — 2003. — Vol. 21, no. 7. — Pp. 818-821.