Математическая биология и биоинформатика. 2012. Т. 7. № 2. С. 554–566.
URL: http://www.matbio.org/2012/Riabenko_7_554.pdf

========== ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ ==========

УДК: 519.254

Настройка нелинейной модели данных экспериментов с экспрессионными ДНК-микрочипами

©2012 Рябенко Е.А.*

Факультет вычислительной математики и кибернетики, Московский государственный университет им. М.В. Ломоносова, Москва, 119991, Россия
НТЦ Биоклиникум, Москва, 115088, Россия

Аннотация. Рассматривается нелинейная модель данных ДНК-микрочипов, в которой интенсивность флуоресценции проб описывается функцией Лэнгмюра. Разработан метод настройки параметров модели на основе общедоступных данных нескольких тысяч экспериментов, основанный на минимизации функции потерь из класса AB-дивергенций; для выбора оптимальных значений гиперпараметров проведены численные эксперименты. Полученная модель описывает интенсивности флуоресценции проб микрочипа точнее стандартной линейной, а полученные на её основе оценки экспрессии более устойчивы.

Ключевые слова: *ДНК-микрочипы, суммаризация, модель Лэнгмюра, AB-дивергенция.*

ВВЕДЕНИЕ

Технология микрочипов ДНК позволяет получить оценку экспрессии десятков тысяч генов одновременно. Основной принцип работы микрочипов ДНК заключается в следующем. На поверхности микрочипа на известных позициях закреплены пробы – одноцепочечные фрагменты ДНК, последовательности нуклеотидов в которых известны. Исследуемый образец специально готовят таким образом, чтобы в нём находились одинарные цепочки ДНК экспрессируемых генов. Согласно принципу комплементарности, одинарные цепочки в образце вступят в реакцию гибридизации с пробами. После этого на образец наносят флуоресцентные метки, чтобы по результатам сканирования микрочипа определить, какие именно участки цепочек ДНК вступили в реакцию, и оценить концентрации соответствующих генов.

В настоящее время существует несколько популярных платформ для микрочипового анализа экспрессии. Технология Affymetrix GeneChip, впервые предложенная в 1996 году, на сегодняшний день является одной из наиболее популярных. В данной работе речь пойдёт о методах анализа данных, полученных при помощи ДНК-микрочипов Affymetrix Human Gene 1.0 ST, относящихся к последнему поколению микрочипов этого производителя.

Для обеспечения устойчивости оценки уровня экспрессии каждому гену на микрочипе соответствует несколько проб; их последовательности комплементарны разным участкам гена. В ходе обработки данных микрочипового анализа на этапе суммаризации интенсивности флуоресценции проб, соответствующих одному гену, обобщаются в оценку его экспрессии. Простейший метод суммаризации – усреднение интенсивностей флуоресценции проб по каждому гену. Такой подход применяется в

* riabenko.e@gmail.com

комплексе алгоритмов MAS 5.0 [1]. Однако основной недостаток такого подхода заключается в том, что разброс интенсивностей различных проб одного гена очень высок (может достигать нескольких порядков), причём эти различия носят систематический характер. Так, в работе [2] было показано, что вариация интенсивностей проб к одному гену на одном микрочипе, как правило, выше, чем вариации интенсивностей этих проб на разных микрочипах; другой пример см. на рис. 1.

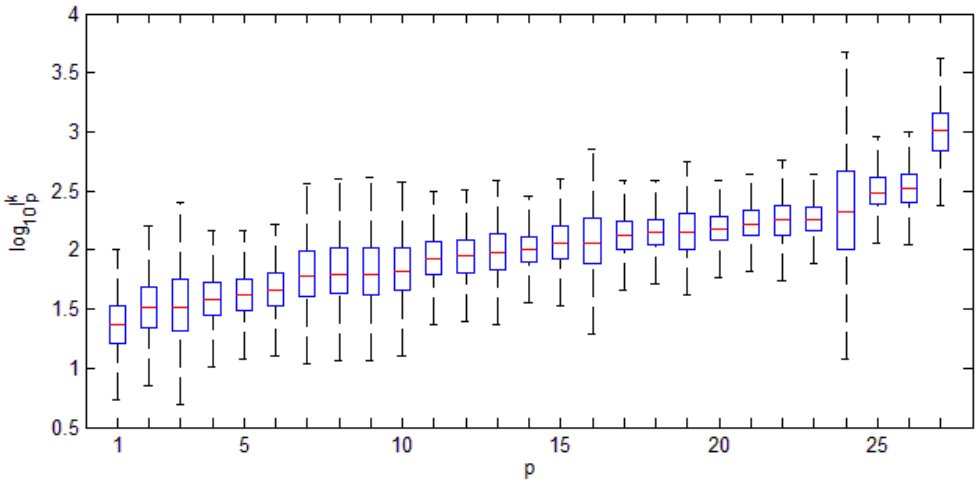

Рис. 1. Распределения интенсивностей флуоресценции разных проб к одному гену на тысяче ДНК-микрочипов: по горизонтальной оси – номер пробы, по вертикальной – десятичный логарифм интенсивности флуоресценции.

Основная причина таких различий заключается в том, что энергия гибридизации пробы с комплементарной ей последовательностью из исследуемого образца существенно зависит от последовательности пробы. Как известно, в двуцепочечной молекуле ДНК тимин (T), как правило, соединяется с аденином (A), а гуанин (G) – с цитозином (C). Пара A/T образует две водородные связи, а пара G/C – три, поэтому энергия взаимодействия тем больше, чем больше в ней пар G/C. Но число A/T и G/C пар – не единственный фактор, определяющий энергию взаимодействия двух цепочек ДНК: она зависит также от местонахождения каждой пары в цепочке, соседних нуклеотидов и т.д. (подробный обзор факторов приведён в работе [3]). В связи с этим на данный момент разработка точной физической модели гибридизации проб на микрочипе не представляется возможной.

Методы суммаризации следующего поколения, такие, как RMA [4], gcRMA [5], dChip [6], PLIER [7], учитывают вариацию энергии гибридизации проб в рамках линейной модели следующего вида:

$$I_p^k = a_p c_{g(p)}^k, \qquad (1)$$

где I_p^k – интенсивность флуоресценции пробы p на микрочипе k, $c_{g(p)}^k$ – уровень экспрессии гена g, которому проба p комплементарна, на микрочипе k, а a_p – коэффициент сродства пробы p своему гену. Основанные на этой модели методы применяются при анализе ДНК-микрочипов практически повсеместно. Однако способ их применения обладает очевидным недостатком. Коэффициенты сродства a_p по определению не зависят от исследуемого в эксперименте образца – они постоянны для каждой фиксированной модели микрочипа. Тем не менее, в наиболее распространённых инструментах анализа коэффициенты сродства каждый раз заново определяются непосредственно по данным анализируемого эксперимента. При этом необходимо,

чтобы выборка микрочипов была достаточно большого размера, а для часто встречающихся на практике небольших выборок оценки экспрессии могут оказаться неустойчивыми. Кроме того, получаемые при независимом анализе разных выборок оценки нельзя непосредственно сравнивать между собой.

Очевидное решение этой проблемы – использование предварительно настроенной модели, в которой значения коэффициентов сродства заранее определены по репрезентативной выборке микрочипов. Такой подход был впервые использован в методе refRMA, предложенном в работе [8]. В предложенной реализации используются коэффициенты сродства, настроенные по выборке из 1614 микрочипов. Как показано авторами, оценки экспрессии, получаемые при помощи данного метода, не уступают результатам применения стандартных методов по ряду критериев качества. К сожалению, область применения программного пакета, реализующего метод refRMA, ограничена микрочипами предыдущего поколения Affymetrix Human Gene U133A.

Дальнейшее развитие данный подход получил в рамках метода frozen RMA (fRMA) в работе [9], где коэффициенты сродства линейной модели были оценены по выборке из 850 микрочипов. В следующей работе [10] авторами был представлен пакет frmaTools, позволяющий провести предварительную настройку модели на собственноручно подобранной выборке микрочипов. Однако для исследуемых микрочипов Affymetrix последнего поколения не подходит и он.

Особенностью публикации результатов, полученных при помощи ДНК-микрочипов, является обязательная загрузка необработанных экспериментальных данных в одну из публично доступных баз данных. Так, наиболее крупная база данных GEO [11] содержит данные об интенсивностях флуоресценции проб нескольких тысяч микрочипов Affymetrix Human Gene 1.0 ST. Таким образом, для настройки параметров модели микрочиповых данных в настоящее время можно использовать выборку достаточно большого объёма, что позволяет надеяться на повышение точности получаемых в результате методов анализа.

Несмотря на то, что абсолютное большинство методов обработки данных микрочиповых экспериментов опирается на предположение о линейной зависимости между интенсивностью флуоресценции пробы и уровнем экспрессии соответствующего гена, неоднократно было показано (см., например, [12]), что эта зависимость лучше описывается нелинейной функцией Лэнгмюра:

$$I_p^k = \frac{a_p c_{g(p)}^k}{1 + b_p c_{g(p)}^k}, \tag{2}$$

где коэффициент $b_p = a_p / I_p^{\max}$ характеризует насыщение, а $I_p^{\max}$ – уровень горизонтальной асимптоты кривой насыщения. Задаваемая функцией Лэнгмюра кривая имеет участок, достаточно хорошо приближаемый прямой, однако в действительности интенсивности флуоресценции проб к генам с высокой экспрессией зачастую оказываются за пределами этого участка [13].

Подобные нелинейные модели микрочиповых данных рассматривались прежде в работах [14–16], однако коэффициенты a_p и b_p определялись не непосредственно по данным, а на основе нуклеотидных последовательностей проб путём приближённого вычисления свободной энергии Гиббса с применением модели ближайшего соседа. В то же время было показано, что свободная энергия Гиббса имеет весьма отдалённое отношение к настоящим коэффициентам сродства и насыщения нелинейной модели, поскольку не учитывает ряд существенных физических факторов [17]. Кроме того, указанные исследования были проведены для микрочипов Affymetrix предыдущего поколения, что делает невозможным их применение к данным экспериментов с микрочипами рассматриваемого нами типа.

Цель данной работы – разработка метода определения параметров модели (2) для микрочипов Affymetrix Human Gene 1.0 ST по данным нескольких тысяч микрочиповых экспериментов. Как показывают вычислительные эксперименты, настроенная модель позволяет точнее объяснять наблюдаемые значения интенсивностей флуоресценции проб, а получаемые на её основе оценки экспрессии обладают большей устойчивостью.

МАТЕРИАЛЫ И МЕТОДЫ

Имеющиеся данные

При помощи пакета GEOquery [18] языка R из базы данных GEO было получено несколько тысяч файлов с интенсивностями флуоресценции проб на микрочипах Affymetrix Human Gene 1.0 ST. Часть данных, содержавших явные ошибки форматирования, была отброшена; для дальнейших экспериментов было отобрано 3459 микрочипов.

В соответствии с аннотацией микрочипа, предоставленной производителем, была проведена фильтрация проб согласно следующим критериям:

– отброшены служебные пробы, флуоресценция которых не является мерой экспрессии какого-либо гена;

– для упрощения задачи исключены пробы, последовательность которых комплементарна сразу нескольким генам, поскольку интенсивность их флуоресценции определяется суммарным уровнем экспрессии нескольких генов (таких проб на рассматриваемом микрочипе ~3.7%);

– наборы проб к одному гену, в которых по результатам предыдущего этапа фильтрации осталось не более трёх проб.

Таким образом, для последующих экспериментов было отобрано 735497 проб к 26902 генам.

Будем использовать следующие обозначения: P – число проб, K – число чипов; G – число генов; $\mathbf{I} \in R_+^{P\times K}$ – интенсивности флуоресценции проб на чипах, $\mathbf{c} \in R_+^{G\times K}$ – уровни экспресии генов, $\mathbf{a},\mathbf{b} \in R_+^{P}$ – коэффициенты сродства и насыщения для проб. Множество проб, соответствующих гену g, обозначим как $P(g)$, а ген, которому соответствует проба p – как $g(p)$.

Этапу суммаризации в обработке микрочиповых данных предшествуют процедуры фоновой поправки и нормализации, позволяющие исключить влияние фонового свечения и выровнять распределения интенсивностей флуоресценции проб на разных микрочипах [19]. В качестве уровня фона для каждого микрочипа была выбрана минимальная интенсивность флуоресценции пробы на нём, и это значение было вычтено из интенсивностей флуоресценции всех проб. Далее была применена медианная нормализация, в ходе которой интенсивности были масштабированы так, чтобы их медианы по каждому микрочипу были равны одному и тому же числу (в данном случае восьмидесяти). Поскольку исследуемые ниже алгоритмы настройки моделей могут быть численно неустойчивыми в случае, когда интенсивности принимают нулевые значения, ко всем интенсивностям была прибавлена единица.

Функция потерь

Для настройки параметров модели необходимо задать некоторую функцию потерь и найти значения, доставляющие ей минимум. Наибольший интерес представляют сепарабельные функции потерь:

$$D(\mathbf{P},\mathbf{Q}) = \sum_{i,j} d(p_{ij}, q_{ij}) \geq 0,$$

где $d(p,q)=0$ тогда и только тогда, когда $p=q$.

Чаще всего в качестве функции потерь используется норма Фробениуса $D(\mathbf{P},\mathbf{Q})=\sum_{i,j}\left(p_{ij}-q_{ij}\right)^2$. Одной из причин такой популярности, наряду с простотой и интуитивной понятностью, является оптимальность получаемых при её минимизации оценок для моделей с аддитивным гауссовским шумом. Однако для других видов шума, а также в присутствии выбросов, оценки, доставляющие минимум норме Фробениуса, могут оказываться несостоятельными. Большей устойчивостью обладают оценки, получаемые при использовании l_1-нормы $D(P,Q)=\sum_{i,j}\left|p_{ij}-q_{ij}\right|$; показана их оптимальность для моделей с аддитивным шумом, имеющим распределение Лапласа.

Для изучения характера распределения шума в интенсивностях флуоресценции проб на микрочипах были использованы предоставленные производителем данные эксперимента [20]. Выборка содержит большое количество так называемых технических репликатов – микрочипов, на которые был нанесён один и тот же материал. Различия в свечении проб на таких чипах объясняется только шумовой компонентой. Распределение попарных разностей интенсивностей флуоресценции проб на технических репликатах показано на рис. 2. Оно имеет крайне тяжёлые хвосты и плохо описывается разностью нормальных или лапласовых случайных величин.

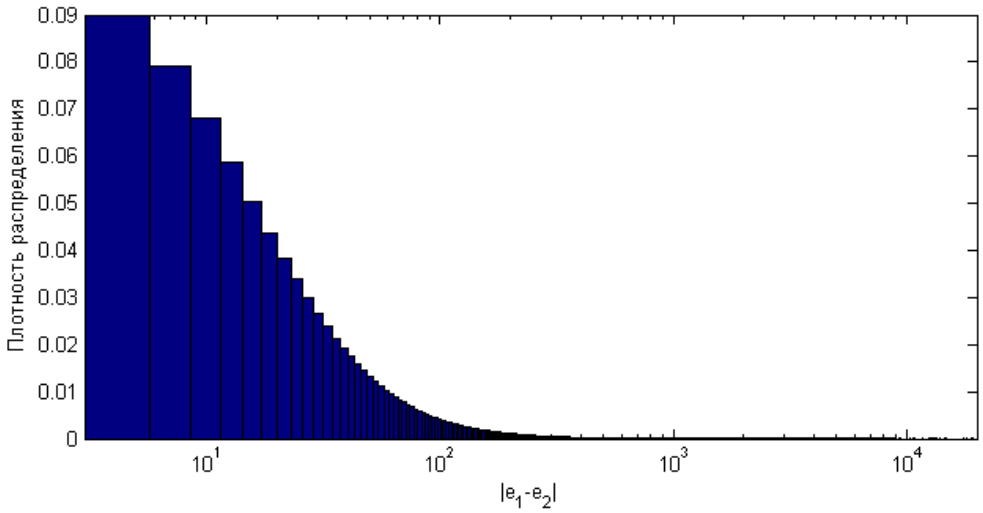

Рис. 2. Гистограмма попарных разностей интенсивностей флуоресценции проб на технических репликатах.

В отсутствие полной информации о характере и виде распределения шума функцию потерь можно выбрать из каких-либо эмпирических соображений. Часто используются робастные M-оценки: так, в методе RMA отклонения моделируемых интенсивностей от реальных измеряются при помощи функции Хубера, в PLIER – функции Гемана-Маклуре.

Другой вариант – взять достаточно широкий параметрический класс функций потерь и выбрать параметры, наилучшие в смысле каких-либо заданных критериев качества. Такими классами являются, например, альфа- и бета-дивергенции, задающие непрерывные по параметрам множества сепарабельных функций потерь, включающих в том числе известные нормы Фробениуса и l_1, дивергенции Кульбака–Лейблера, Итакура–Саито, расстояния Хелинджера, хи-квадрат Пирсона и Неймана и многие другие. В данной работе было решено использовать предложенный в [21] класс AB-дивергенций, обобщающий альфа- и бета-дивергенции и задаваемый в виде двухпараметрического семейства функций потерь следующего вида:

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

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

AB-дивергенции – одно из наиболее широких используемых на сегодняшний день семейств функций потерь. Оценки, получаемые при минимизации функций этого класса, являются оценками максимального правдоподобия в условиях самых разнообразных видов распределений шума, как аддитивного и мультипликативного, так и смешанного, состоящего из обеих этих компонент. Выбирая по данным параметры α и β, показывающие наилучшие значения критериев качества, мы тем самым неявно выбираем модель шума, оптимальную для исследуемой модели данных в смысле этих критериев.

Критерии качества

Для оценки качества построенных моделей использовались следующие критерии.

Точность приближения. Перед тем, как формализовать данный критерий, обратим внимание ещё на одну особенность структуры исследуемых данных. Интенсивности флуоресценции проб к одному гену могут меняться несогласованно (см. пример на рис. 3). Дело в том, что пробы, имеющиеся на микрочипе, комплементарны участкам, расположенным по всей длине гена, а из-за явления, известного как альтернативный сплайсинг, некоторые участки одноцепочечной ДНК в наносимой на микрочип смеси могут отсутствовать. Интенсивность флуоресценции пробы *p*, соответствующей такому участку, может оказаться низкой даже тогда, когда уровень экспрессии гена высок. В этом случае модельная интенсивность $\hat{I}_p^k = a_p c_{g(p)}^k \big/ (1+b_p c_{g(p)}^k)$ будет выше фактической интенсивности I_p^k.

С учётом этого наблюдения зададим критерий точности приближения в следующем виде:

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

Здесь **W** – бинарная матрица весов, ноль в которой соответствует пробе, флуоресценция которой искажена в результате альтернативного сплайсинга. Веса оцениваются при помощи итеративной процедуры, описанной ниже.

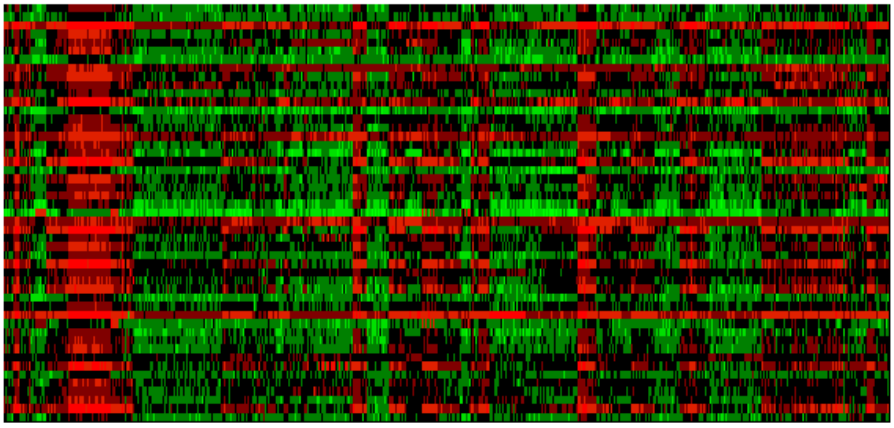

Рис. 3. Тепловая карта логарифмов интенсивностей флуоресценции проб к одному гену на тысяче ДНК-микрочипов: по горизонтальной оси – микрочипы, по вертикальной – пробы.

Воспроизводимость параметров пробы. Разобьём исходную выборку микрочипов на две равные части, по каждой из них восстановим векторы коэффициентов **a** и **b**. Чем ближе друг к другу полученные значения, тем лучше. Рассмотрим величины

$$\frac{|a_{1p}-a_{2p}|}{a_{1p}+a_{2p}}, \qquad \frac{|b_{1p}-b_{2p}|}{b_{1p}+b_{2p}}.$$

Они принимают значения на отрезке [0,1], и эксперименты показывают, что каждая из них имеет распределение с двумя выраженными пиками в нуле и единице, поэтому усреднять по p нецелесообразно. Исходя из этого, зададим критерии воспроизводимости параметров пробы в следующем виде:

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

Квадратные скобки здесь означают индикаторную функцию (нотация Айверсона).

Воспроизводимость оценок экспрессии. Для гена g исключим из рассмотрения одну пробу p_i из имеющегося набора $P(g)$; обозначим за $\mathbf{c}_{g,\bar{p}_i}$ вектор оценок экспрессии этого гена, построенный по множеству проб $P(g)\setminus p_i$. Мерой воспроизводимости в таком случае будет расстояние между $\mathbf{c}_g$ и $\mathbf{c}_{g,\bar{p}_i}$. Для большей устойчивости будем повторять процедуру исключения для пяти разных проб. Таким образом, критерий воспроизводимости оценок экспрессии задаётся следующим образом:

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

Оптимизационная задача

Параметры предложенной нелинейной модели находятся как решение следующей оптимизационной задачи:

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

$$a_p \ge 0, \quad b_p \ge 0, \quad p=1,\ldots,P, \qquad c_g^k \ge 0, \quad k=1,\ldots,K, \quad g=1,\ldots,G.$$

Благодаря сепарабельности функции потерь оптимизационная задача распадается на G независимых подзадач по каждому гену g:

$$\sum_{k=1}^{K}\sum_{p\in P(g)} d_{AB}^{(\alpha,\beta)}\left(I_p^k, \frac{a_p c_g^k}{1+b_p c_g^k}\right)\to \min_{\mathbf{a}_g,\mathbf{b}_g,\mathbf{c}_g},$$
$$a_p \ge 0, \quad b_p \ge 0, \quad p\in P(g),$$
$$c_g^k \ge 0, \quad k=1,\ldots,K.$$

Пусть набор векторов $\mathbf{a}_g,\mathbf{b}_g,\mathbf{c}_g$ – решение приведённой задачи, тогда векторы $(1/C)\cdot\mathbf{a}_g,(1/C)\cdot\mathbf{b}_g,C\cdot\mathbf{c}_g$ тоже будут являться решением, то есть, коэффициенты модели находятся с точностью до константного множителя. Для однозначности определения коэффициентов добавим условие нормировки $\prod_{p\in P(g)} a_p = 1$, используемое в аналогичных моделях микрочиповых данных (RMA, PLIER). Кроме того, для обеспечения устойчивости получаемых оценок экспрессии добавим квадратичную регуляризацию по $\mathbf{c}$ с параметром α_c. С учётом этих дополнений для каждого гена g оптимизационная задача имеет следующий вид:

$$f\left(\mathbf{I}_g,\mathbf{a}_g,\mathbf{b}_g,\mathbf{c}_g,\alpha_c\right)=\sum_{k=1}^{K}\sum_{p\in P(g)} d_{AB}^{(\alpha,\beta)}\left(I_p^k, \frac{a_p c_g^k}{1+b_p c_g^k}\right)+\frac{\alpha_c}{2}\sum_{k=1}^{K}\left(c_g^k\right)^2\to \min_{\mathbf{a}_g,\mathbf{b}_g,\mathbf{c}_g},$$
$$\prod_{p\in P(g)} a_p = 1, \qquad (3)$$
$$a_p \ge 0, \quad b_p \ge 0, \quad p\in P(g),$$
$$c_g^k \ge 0, \quad k=1,\ldots,K.$$

Далее для простоты индекс g будем опускать.

Для решения оптимизационной задачи (3) будем использовать метод блочно-покоординатного спуска, делая шаги стандартного метода Ньютона с проекцией на положительную область поочерёдно по $\mathbf{a}$, $\mathbf{b}$ и $\mathbf{c}$. При этом сепарабельность функции потерь приводит к тому, что каждая из трёх задач минимизации распадается на независимые одномерные задачи:

$$a_p = \max\left(0, a_p - \frac{\partial f\left(I,a,b,c,\alpha_c\right)}{\partial a_p}\Big/\frac{\partial^2 f\left(I,a,b,c,\alpha_c\right)}{\partial a_p^2}\right), \quad p\in P(g),$$
$$c^k = \max\left(0, c^k - \frac{\partial f\left(I,a,b,c,\alpha_c\right)}{\partial c^k}\Big/\frac{\partial^2 f\left(I,a,b,c,\alpha_c\right)}{\partial c^{k2}}\right), \quad k=1,...,K,$$
$$b_p = \max\left(0, b_p - \frac{\partial f\left(I,a,b,c,\alpha_c\right)}{\partial b_p}\Big/\frac{\partial^2 f\left(I,a,b,c,\alpha_c\right)}{\partial b_p^2}\right), \quad p\in P(g).$$

Производные здесь могут быть найдены аналитически.

Эксперименты показали, что сходимость процесса оптимизации существенным образом зависит от начального приближения (рис. 4). При этом если в качестве начального приближения для $\mathbf{a}$ и $\mathbf{c}$ брать векторы $\mathbf{a}^L,\mathbf{c}^L$, полученные как результат минимизации той же AB-дивергенции относительно линейной модели интенсивностей $I_p^k = a_p c_{g(p)}^k$, а начальное приближение для $\mathbf{b}$ брать равным нулю, сходимость в основной локальный минимум происходит достаточно быстро. Настройка параметров

линейной модели проводилась при помощи блочно-покоординатного мультипликативного алгоритма, предложенного в [21].

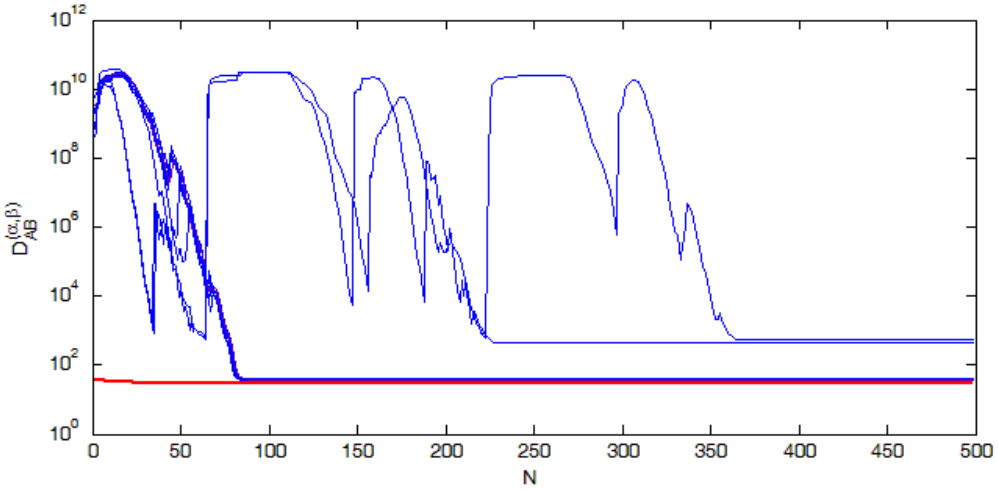

Рис. 4. Зависимость функции потерь от номера итерации при настройке нелинейной модели со случайной инициализацией (синие линии) и с инициализацией $\mathbf{a}^0 = \mathbf{a}^L, \mathbf{b}^0 = \mathbf{0}, \mathbf{c}^0 = \mathbf{c}^L$ (красная линия).

Для уменьшения воздействия на модель проб с интенсивностью, низкой из-за альтернативного сплайсинга, была построена следующая итеративная процедура.

Настроив нелинейную модель, рассчитаем ошибку приближения интенсивностей флуоресценции с весами, пропорциональными концентрациям и обратно пропорциональными фактическим интенсивностям:

$$E_p^k = \frac{\dfrac{a_p c_{g(p)}^k}{1 + b_p c_{g(p)}^k} - I_p^k}{I_p^k} \cdot c_{g(p)}^k. \qquad (4)$$

Для проб, затронутых эффектом альтернативного сплайсинга, значение ошибки будет большим, тем больше, чем больше экспрессия гена и чем меньше интенсивность флуоресценции пробы. Отберём 5% проб, дающих наибольшую ошибку E_p^k, и создадим матрицу весов $\mathbf{W} \in R^{P \times K}$, заполненную следующим образом:

$$W_p^k = \begin{cases} 0, & E_p^k \geq q_{0.95}, \\ 1 & \text{иначе.} \end{cases}$$

Здесь $q_{0.95}$ – 95% выборочный квантиль E_p^k. На следующем шаге итерационного процесса при обновлении **a**, **b** и **c** будем учитывать только те компоненты производных, которые имеют в матрице **W** ненулевые веса.

В организованном таким образом процессе на каждом шаге исключается из рассмотрения 5% проб; эксперименты показывают, что в среднем после исключения 25% проб дальнейшие итерации не приводят к значительным изменениям модели.

Необходимо также каким-то образом выбрать значения коэффициента регуляризации α_c. Учитывая, что характерный диапазон значений AB-дивергенций существенно меняется вместе с параметрами α и β, подбирать оптимальное значение коэффициента необходимо отдельно для каждой пары (α, β). Для фиксированного значения α_c после настройки на обучающей выборке векторов параметров проб

$\mathbf{a}_{train}, \mathbf{b}_{train}$ описанным выше итерационным методом на валидационной выборке вычисляется точность приближения $fit\left(\mathbf{I}_{val}, \mathbf{a}_{train}, \mathbf{b}_{train}, \mathbf{c}_{val}, \alpha_c\right)$. В качестве оптимального значения α_c^{best} выбирается $\arg\min_{\alpha_c} fit\left(\mathbf{I}_{val}, \mathbf{a}_{train}, \mathbf{b}_{train}, \mathbf{c}_{val}, \alpha_c\right)$, а для минимизации этой функции используется метод золотого сечения (см., например, [22]).

Результаты экспериментов

Для настройки модели и расчёта значений критериев качества были сформированы обучающая, валидационная и тестовая выборки размером в 200 микрочипов каждая. Решение задачи (3) было получено для $(\alpha, \beta) \in [-2,4] \times [-4,4]$ с шагом 0.5. Значения критериев качества отображены на рис. 5.

В областях, обозначенных белым, значения критериев качества получились значительно выше (для точности приближения – на порядки). По всей видимости, это связано с тем, что в этих областях AB-дивергенции не являются выпуклыми функциями [21], из-за чего процесс минимизации (3) не сходится.

По совокупности рассмотренных критериев качества наилучший результат был получен при значениях параметров $\alpha = 2, \beta = 1$.

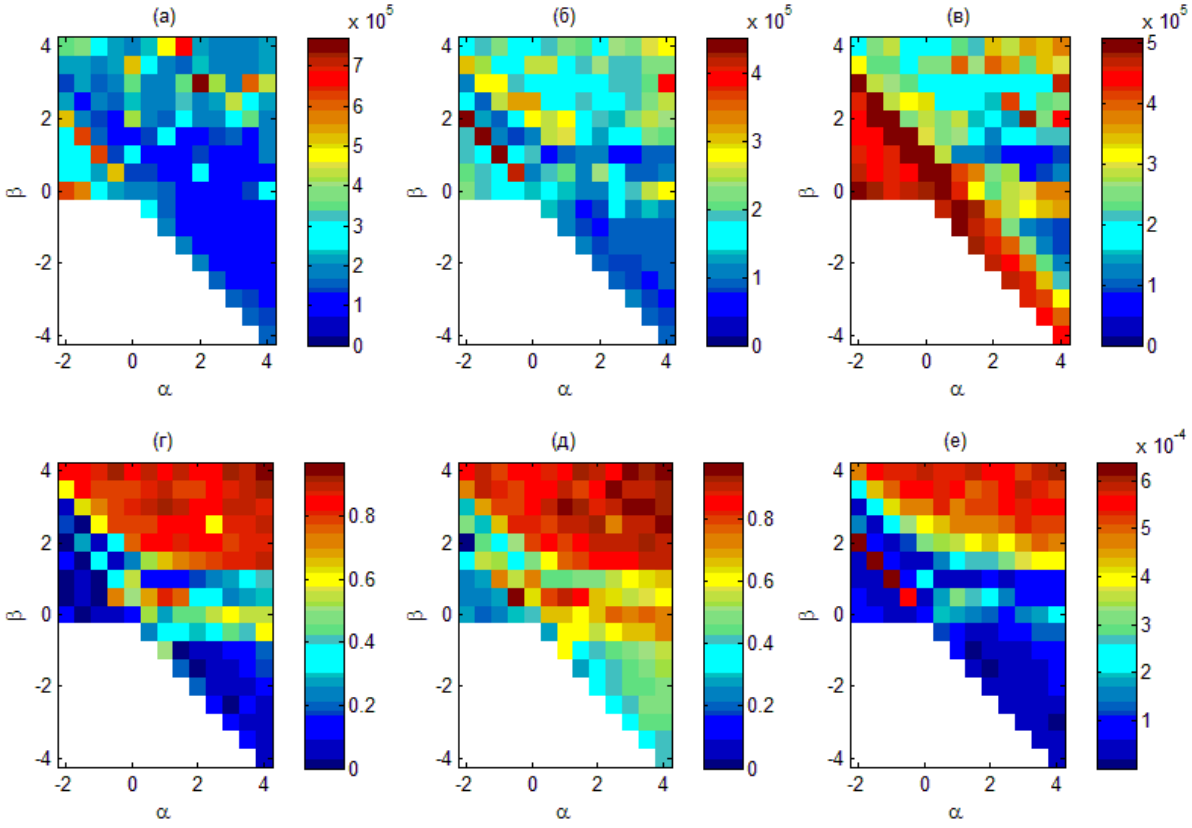

Рис. 5. Значения критериев качества в зависимости от α и β. Верхний ряд – точность приближения: (а) – на обучающей выборке, (б) – на валидационной, (в) – на тестовой; нижний ряд – воспроизводимость: (г) – коэффициентов сродства, (д) – коэффициентов насыщения, (е) – оценок экспрессии.

Для сравнения с выбранным алгоритмом значения критериев качества были рассчитаны для метода RMA, являющегося на сегодняшний день де-факто стандартом в анализе микрочипов, а также для его модификации, отбрасывающей по аналогии с предложенным алгоритмом 25% проб с высокими значениями ошибки (4). Полученные

значения приведены в таблице 1. Так как в основе RMA лежит линейная модель, величина rep_b для этого метода не определена.

Таблица 1. Значения критериев качества в проведённых экспериментах

	Предлагаемый алгоритм	RMA	RMA с фильтрацией
$fit(I_{train})$	7.441×10^{3}	3.087×10^{4}	2.028×10^{4}
$fit(I_{test})$	7.742×10^{3}	4.080×10^{4}	2.680×10^{4}
rep_a	0.0772	2.395×10^{-4}	1.573×10^{-3}
rep_b	0.3667	–	–
rep_c	0.0395	0.1495	0.0956

Для выбранных значений α, β была проведена повторная настройка модели с использованием обучающих выборок большего размера. Зависимость точности приближения от размера обучающей выборки показана на рис. 6.

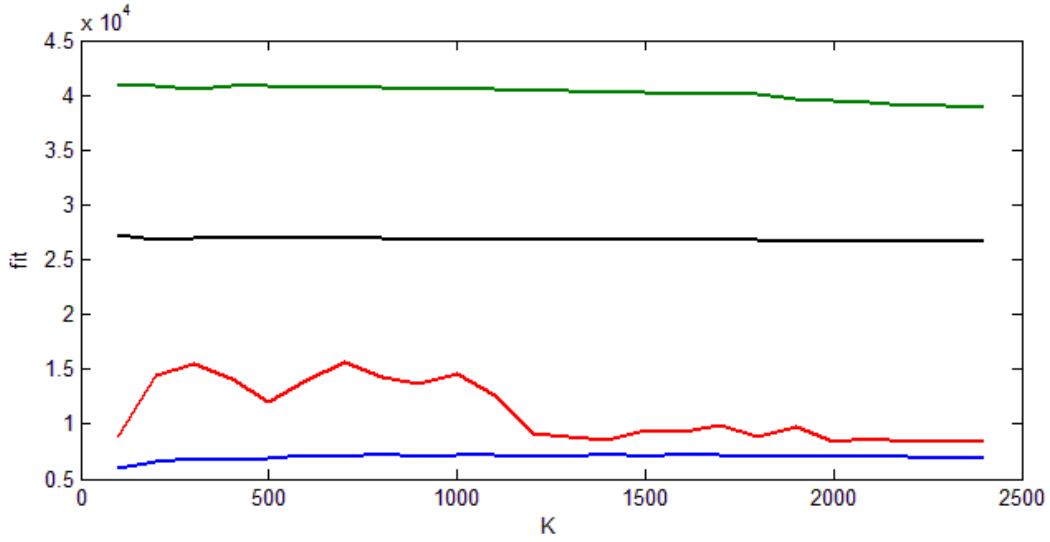

Рис. 6. Зависимость точности приближения от размера обучающей выборки: синий – построенная модель, обучающая выборка; красный – построенная модель, тестовая выборка; зелёный – линейная модель, метод RMA, тестовая выборка; чёрный – линейная модель, метод RMA с отбрасыванием 25% проб с высокими значениями ошибки (4), тестовая выборка.

По результатам моделирования можно сделать следующие выводы. При настройке нелинейной модели интенсивности флуоресценции проб предложенным алгоритмом точность приближения получается выше, чем при использовании линейной модели и метода RMA. Учёт эффекта альтернативного сплайсинга при помощи исключения из рассмотрения части проб позволяет увеличить точность приближения линейной модели (для нелинейной модели наблюдался аналогичный эффект). Метод RMA даёт более устойчивые оценки коэффициентов сродства, однако по воспроизводимости оценок экспрессии предложенный метод показывает лучшие результаты.

ЗАКЛЮЧЕНИЕ

В данной работе рассматривалась нелинейная модель интенсивности флуоресценции проб в экспериментах с экспрессионными ДНК-микрочипами

Affymetrix Human Gene 1.0 ST. Для настройки параметров модели была использованы данные нескольких тысяч микрочиповых экспериментов, полученные из общедоступной базы данных GEO. При разработке метода настройки модели минимизируемая функция потерь была выбрана в классе AB-дивергенций согласно заданным критериям качества. Полученная модель точнее описывает интенсивности флуоресценции проб, чем модель, получаемая наиболее распространённым стандартным методом (RMA), а созданный на её основе метод оценивания экспрессии обладает большей устойчивостью.

Предложенный метод планируется оформить в виде пакета для языка R и предоставить в общее пользование средствами платформы Bioconductor.

Работа выполнена при поддержке Министерства образования и науки (ГК № 16.522.11.2004) и гранта РФФИ № 12-07-31200/12.

СПИСОК ЛИТЕРАТУРЫ

1. Hubbell E., Liu W.-M., Mei R. Robust estimators for expression analysis. *Bioinformatics*. 2002. V. 18. № 12. P. 1585–1592.
2. Li C, Wong W.H. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*. 2001. V. 2. № 8. RESEARCH0032.
3. Koltai H., Weingarten-Baror C. Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction. *Nucleic acids research*. 2008. V. 36. № 7. P. 2395–405.
4. Irizarry R.A., Hobbs B., Collin F., Beazer-Barclay Y.D., Antonellis K.J., Scherf U., Speed T.P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003. V. 4. № 2. P. 249–264.
5. Wu Z., Irizarry R.A., Gentleman R., Martinez-Murillo F., Spencer F. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*. 2004. V. 99. № 468. P. 909–917.
6. 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. V. 98. № 1. P. 31–36.
7. Affymetrix. *Guide to Probe Logarithmic Intensity Error (PLIER) Estimation. Technical note*. URL: http://media.affymetrix.com/support/technical/technotes/plier_technote.pdf (дата обращения: 02.09.2012).
8. Katz S., Irizarry R.A., Lin X., Tripputi M., Porter M.W. A summarization approach for Affymetrix GeneChip data using a reference training set from a large, biologically diverse database. *BMC bioinformatics*. 2006. V. 7. P. 464.
9. McCall M.N., Bolstad B.M., Irizarry R.A. Frozen robust multiarray analysis (fRMA). *Biostatistics*. 2010. V. 11. № 2. P. 242–253.
10. McCall M.N., Irizarry R.A. Thawing Frozen Robust Multi-array Analysis (fRMA). *BMC Bioinformatics*. 2011. V. 12. № 1. P. 369.
11. The National Center for Biotechnology Information. *Gene Expression Omnibus*. URL: http://www.ncbi.nlm.nih.gov/geo/ (дата обращения: 02.09.2012).
12. Halperin A., Buhot A. Zhulina E.B. Sensitivity, specificity, and the hybridization isotherms of DNA chips. *Biophysical Journal*. 2004. V. 86. P. 718–730.
13. Abdueva D., Skvortsov D., Tavaré S. Non-linear analysis of GeneChip arrays. *Nucleic acids research*. 2006. V. 34. № 15. P. e105.
14. Heim T., Wolterink J.K., Carlon E., Barkema G.T. Effective affinities in microarray data. *Journal of Physics: Condensed Matter*. 2006. V. 18. P. S525.

15. Held G.A., Grinstein G., Tu Y. Relationship between gene expression and observed intensities in DNA microarrays – a modeling study. *Nucleic acids research*. 2006. V. 34. № 9. P. e70.
16. Binder H., Preibisch S., Berger H. Calibration of microarray gene-expression data. *Methods in Molecular Biology*. 2010. V. 576. № 16. P. 375–407.
17. Krainer S. *Microarray based real-time analysis of nucleic acid hybridization kinetics and thermodynamics*. Dissertation, Dept. of Biology, Johannes Gutenberg-Universitat Mainz, 2011.
18. Davis S., Meltzer P.S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*. 2007. V. 14. P. 1846–1847.
19. Huber W, Irizarry R.A., Gentleman R. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Eds. Gentleman R., Carey V.J., Huber W., Irizarry R.A., Dudoit S. New York: Springer Science+Business Media, 2005. P. 3–12.
20. Affymetrix. *Sample Data, Gene 1.0 ST Data Set*. URL: http://www.affymetrix.com/support/technical/sample_data/gene_1_0_array_data.affx (дата обращения: 02.09.2012).
21. Cichocki A., Cruces S., Amari S. Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization. *Entropy*. 2011. V. 13. № 1. P. 134–170.
22. Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P. *Numerical Recipes: The Art of Scientific Computing (3rd ed.)*. New York: Cambridge University Press, 2007. 1256 p.

Материал поступил в редакцию 02.10.2012, опубликован 25.10.2012.