Биотехнологии

УДК 519.2:577.2

Экспрессия генов и микрочипы: проблемы количественного анализа

А. Н. Свешникова, П. С. Иванов

АНАСТАСИЯ НИКИТИЧНА СВЕШНИКОВА– аспирантка кафедры биофизики Физического факультета МГУ им. М.В.Ломоносова. Область научных интересов: вычислительная геномика, математические методы анализа функционирования генома.

ПАВЕЛ СЕРГЕЕВИЧ ИВАНОВ– кандидат физико-математических наук, старший научный сотрудник кафедры биофизики Физического факультета МГУ. Область научных интересов: математические методы анализа функционирования генома, методы нелинейной динамики в исследовании биологических ритмов.

119992 Москва, Ленинские горы, МГУ им. М.В. Ломоносова, Физический факультет, тел. (495)939-3025, факс (495)932-8820, E-mail p-ivanov@mtu-net.ru

Технология генных микрочипов, совершившая настоящую революцию в области исследования генетического аппарата [1], оказала влияние на развитие многих медико-биологических дисциплин, включая онкологию [2,3], токсикологию [4], фармакологию [5], биологию развития [6], способствовала решению классических молекулярно-биологических задач, связанных с изучением генных мутаций и механизмов транскрипции [7,8]. В отличие от традиционных подходов в молекулярной генетике, таких как методы, основанные на полимеразной цепной реакции (ПЦР), Nothern-блоттинг и последовательный анализ генной экспрессии [9—11], нацеленные на изучение экспрессии одиночных либо небольшого числа генов, микрочипы сделали возможным одновременный мониторинг экспрессии всего клеточного генома при строго контролируемых условиях. Такой мониторинг позволяет выявить стратегии, используемые клеткой в ответ на изменение внешних условий, восстановить группы генов, функционально связанные друг с другом, реконструировать механизмы регуляции транскрипции и определить связанные с ними метаболические пути, проаннотировать гены, функция которых в клетке оставалась неизвестной.

Реализация этой программы требует корректной содержательной интерпретации экспериментального материала, получаемого с применением генных микрочипов. В случае клеток млекопитающих обработка данных, выдаваемых одним микрочипом, порождает десятки тысяч значений экспрессии. Количественный анализ подобных массивов данных составляет предмет самостоятельной области биоинформатики и вычислительной геномики.

В данном обзоре обсуждаются математико-статистические методы, нашедшие применение в исследованиях профилей генной экспрессии с помощью микрочипов. Для удобства рассмотрения методы сгруппированы в соответствии с классами решаемых задач. Особое внимание уделено алгоритмам, специально разработанным для анализа результатов микрочиповых экспериментов.

Технология приготовления и использования генных микрочипов

Новая эра в исследовании генной экспрессии началась с работы [12], авторы которой предложили оценивать уровень экспрессии методом гибридизации молекул ДНК с нуклеотидными последовательностями-зондами, локализованными в определенном участке микрочипа.

Техника экспериментального исследования экспрессии генома посредством микрочипов состоит в следующем (см. рисунок). Выделенная из клетки матричная РНК (мРНК) при помощи обратной транскриптазы превращается в однонитевую клонированную ДНК (ДНК-мишень). Количество синтезируемых молекул ДНК-мишеней пропорционально уровню экспрессии соответствующего гена при заданных условиях [13].

На стадии обратной транскрипции в синтезируемую ДНК-мишень встраиваются радиоактивные или флуоресцентные метки. В первом случае используются молекулы дезоксицитидинтрифосфата, меченные слаборадиоактивным изотопом ^{33}P (^{33}P-дЦТФ). В качестве флуоресцентных меток наибольшее распространение получили водорастворимые цианиновые красители Cy3-дУТФ и Cy5-дУТФ, являющиеся производными 5-(3-амино)аллилдезоксиуридин-5′-фосфата. Указанные молекулы отличаются высокой эффективностью встраивания (благодаря наличию гидроксисукцинимидной группы), хорошей фотостабильностью и высоким квантовым выходом люминесценции. Меченые ДНК-мишени затем гибридизуются с последовательностями-зондами в различных участках микрочипа. При использовании флуоресцентных меток процедура сканирования микрочипа включает в себя возбуждение красителей под действием лазерного излучения вблизи их максимумов поглощения (обычно на длинах волн 532 нм для Cy3-дУТФ и 635 нм для

Основные этапы экспериментального исследования экспрессии генов с использованием микрочипов

Cy5-дУТФ) и последующее измерение интенсивности флуоресценции в диапазонах 550—600 нм (Cy3-дУТФ) и 655—695 нм (Cy5-дУТФ). Поскольку точное расположение последовательностей-зондов фиксировано и известно заранее, относительные уровни экспрессии разных генов могут быть однозначно восстановлены по интенсивности сигналов, полученных с разных участков микрочипа, и затем откалиброваны с учетом фоновой интенсивности сигналов с данного микрочипа. В олигонуклеотидных микрочипах производства фирмы Affymetrix гибридизации могут подвергаться сами молекулы клеточной мРНК, меченные биотином. В этом случае степень гибридизации определяется при помощи стрептавидин-фикоэритриновых конъюгатов, которые дают флуоресцеиновый (зеленый) и фикоэритриновый (красный) гибридизационные сигналы, регистрируемые конфокальным сканирующим микроскопом [14].

По типу последовательностей-зондов микрочипы делятся на два основных класса: олигонуклеотидные микрочипы и микрочипы на основе кодирующей ДНК (кДНК). Первые представляют собой квадратные кварцевые пластинки с рабочей площадью 1,5—2 см^2, к которым пришиваются или на которых синтезируются олигонуклеотиды длиной 25—80 оснований, комплементарные соответствующим участкам ДНК-мишени [15]. Существует несколько технологий изготовления олигонуклеотидных микрочипов: струйное нанесение в сочетании с амидофосфатным синтезом олигонуклеотидов [14], пришивание заранее синтезированных нуклеотидных последовательностей на пластинки, покрытые гелем из разветвленных полимерных молекул [16], комбинаторный синтез с применением фотолитографических масок [17]. Для повышения селективности гибридизации наряду с «правильными» олигонуклеотидами на чип помещаются их копии с одним замещенным основанием. Олигонуклеотидный микрочип обычно содержит 11—20 таких пар оснований для каждого гена и до 20 тысяч различных последовательностей-зондов.

Микрочипы на основе кДНК содержат последовательности-зонды, в роли которых выступают ПЦР-продукты целых молекул кДНК либо известных экспрессируемых генных фрагментов длиной 500—2000 пар оснований [13, 18]. Отбор зондов, наносимых на микрочип, зависит от целей исследования и представляет собой нетривиальную задачу из-за необходимости достижения баланса между чувствительностью и избирательностью связывания ДНК-мишени с зондом. Проблема избирательности особенно актуальна при наличии в исследуемом образце нескольких высокогомологичных генов, например кодирующих консервативную субъединицу в семействе родственных белков [19]. Для повышения селективности в последовательность-зонд иногда включают высокоспецифичные участки генов с нетранслируемым 3′-концом молекулы ДНК, хотя это и снижает чувствительность метода.

В качестве подложки в микрочипах на основе кДНК используются нитроцеллюлозные или заряженные нейлоновые мембраны, а также стеклянные пластинки, обладающие низкой собственной флуоресценцией. Для повышения гидрофобности и адсорбционной способности их покрывают аминосиланами или полилизиновыми олигомерами [18]. Главные достоинства микрочипов на основе мембран — возможность многократного использования и приспособленность для работы с радиоактивными метками, которые обеспечивают большую чувствительность. Преимуществами стеклянных подложек являются возможность ковалентного связывания молекул ДНК, небольшой рабочий объем, в котором происходит гибридизация, устойчивость к воздействию повышенных температур и растворов с высокой ионной силой.

Микрочипы со стеклянной подложкой обычно применяются для сравнения экспрессий генов из двух образцов или значений экспрессии при двух различных условиях. Для этого исследуемые молекулы ДНК помечают разными красителями, после чего осуществляют одновременную их гибридизацию на чипе. При работе с мембранными микрочипами реакции гибридизации проводятся последовательно или параллельно.

Оба вида генных микрочипов имеют сильные и слабые стороны. Использование олигонуклеотидов позволяет обойтись без бактериальных клонов и продуктов ПЦР, различить высокогомологичные гены из одного семейства, учесть эффекты кросс-гибридизации и отфильтровать случаи неспецифического связывания. Вместе с тем применение олигонуклеотидных микрочипов ограничивается тем, что они довольно дороги и выпускаются только крупными компаниями, для считывания данных с них требуется специальное оборудование, набор олигонуклеотидов зачастую содержит до 25% ошибочных последовательностей [20]. В этом отношении микрочипы на базе кДНК имеют явные преимущества: их изготовление не сопряжено со значительными затратами и возможно в лабораторных условиях, считывание полученных результатов не требует высокоспециализированных устройств, а применение длинных фрагментов кДНК обеспечивает высокую специфичность их связывания с ДНК-мишенями. В то же время эти микрочипы не обладают достаточной селективностью для различения близких гомологов генов, технология их приготовления связана с трудоемкими процедурами синтеза,

очистки и хранения молекул кДНК перед их помещением на чип, требуются значительные количества клеточной РНК для достижения приемлемого соотношения сигнал/шум [13]. Но главный их недостаток — невысокая точность и худшая (по сравнению с олигонуклеотидными микрочипами) воспроизводимость результатов [20]. Последнее обстоятельство самым негативным образом сказывается на количественном анализе экспериментальных данных.

Первичная обработка данных микрочиповых экспериментов

Известное расположение последовательностей-зондов на микрочипе позволяет пересчитать интенсивности излучения от изотопных или флуоресцентных меток в относительные уровни экспрессии отдельных генов. Эта процедура — нормировка, а также восстановление пропущенных значений и фильтрация выбросов относятся к первичной обработке данных, цель которой — сделать поправку на фоновую интенсивность излучения и устранить систематические погрешности, которые проявляются при работе с несколькими микрочипами.

Нормировка

Предложено множество методов первичной обработки результатов микрочиповых экспериментов [21—23], и проблема выбора конкретного алгоритма возникает уже на стадии нормировки — преобразования интенсивностей флуоресценции или радиоактивного распада в значения экспрессии. Это преобразование может быть представлено функцией $y_{ij} = f(x_{ij})$, которая связывает уровень экспрессии i-го гена y_{ij}, измеренный на j-м микрочипе, с его истинным значением x_{ij} при данных экспериментальных условиях.

В первых программах анализа результатов изучения экспрессии генов с применением олигонуклеотидных микрочипов в качестве значения экспрессии выбиралась усредненная разность уровней излучения от меток, связанных с «правильными» (PM_i) и модифицированными (MM_i) олигонуклеотидами. На практике это приводило к тому, что для 30—40% генов нормированные значения экспрессии $PM_{ij} - MM_{ij}$ оказывались отрицательными, т.е. гибридизация мишени с измененным олигонуклеотидом происходила интенсивнее, чем с неизмененным, что указывало на ошибки в исходном наборе олигонуклеотидов, помещаемых на чип. Первая линейная модель для разности $PM_{ij} - MM_{ij}$, позволившая учесть систематические погрешности, была предложена в работе [24].

В настоящее время общепринятой процедурой перехода от «изображения» микрочипа к значениям экспрессии является алгоритм Robust Multi-chip Analysis (RMA), оперирующий только значениями PM_{ij}, но учитывающий фоновую интенсивность BG флуоресценции или радиоактивного излучения [25]. Помимо нормировки величин PM_{ij} на уровень шума, в RMA введена процедура квантильной нормировки разности ($PM_{ij} - BG$) по нескольким микрочипам. Полное игнорирование величин MM_{ij} обусловлено тем, что их учет практически всегда приводит к повышению уровня шума. Тем не менее уже первые работы, в которых проводилось сравнение различных методов нормировки с использованием реальных экспериментальных данных, подтвердили достаточно высокую точность, чувствительность и избирательность алгоритма RMA [26].

Процедуры нормировки при работе с микрочипами на базе кДНК нацелены на идентификацию и устранение систематических погрешностей, обусловленных различиями в концентрации и интенсивности флуоресценции применяемых меток-красителей, а также в исходном количестве мРНК [22]. Для определения введенной выше функции $f(x_{ij})$ используются различные аппроксимации. Простейший алгоритм нормировки на суммарную интенсивность флуоресценции предполагает линейную зависимость между величинами y_{ij} и x_{ij} [27]. Метод нормировки с дополнительной калибровкой основан на допущении, что лишь небольшая часть генов существенно изменяет свой уровень экспрессии [28]. Это также позволяет ограничиться линейной функцией $y_{ij} = (x_{ij} - \alpha)/\beta$, где α и β — некоторые коэффициенты, хотя тот же метод позволяет учесть и нелинейные эффекты [23].

Параметры функции $f(x_{ij})$ определяются путем анализа сигналов от всего микрочипа либо от небольшой группы генов, экспрессия которых остается практически постоянной. В сравнительных тестах с использованием двух красителей для двух экспериментальных условий нормировка на постоянный уровень экспрессии включает в себя независимое ранжирование интенсивностей флуоресценции для каждого из красителей Cy3 и Cy5, отбрасывание крайних значений и выявление генов, получивших близкие ранги относительно обоих красителей [29]. Однако даже после такой нормировки в массиве значений может сохраняться асимметрия их распределения, связанная с эффектами красителей. Она наглядно проявляется на графике зависимости логарифма отношения интенсивностей свечения двух красителей $\log_2$(Cy5/Cy3) от логарифма средней интенсивности $\log_2$(Cy5· Cy3) [30]. Процедура коррекции асимметрии предложена в [29]. Ряд других методов нормировки и результаты сравнительного анализа различных алгоритмов описаны в [21, 23, 30].

Перечисленные методы нормировки могут быть усовершенствованы путем разделения микрочипа на области, для каждой из которых параметры функции $f(x_{ij})$ рассчитываются независимо [23]. Если каждый эксперимент проводится несколько раз, то общий подход к нормировке состоит в выборе одного из чипов (реплики) в качестве базового и в применении процедуры нормировки к остальным микрочипам [22, 31]. С другой стороны, когда одна и та же ДНК-мишень гибридизуется на нескольких микрочипах, изменения в количестве клеточной мРНК или красителя могут рассматриваться как часть экспериментальной погрешности. Это приводит к простейшей линейной модели $y_{ijk} = x_{ij} + \varepsilon_{ijk}$, где k — номер реплики, а связанная с ней ошибка ε_{ijk} подчиняется нормальному распределению $N(0,\sigma^2)$ [23].

Более строгое моделирование систематических погрешностей рассмотрено в [32].

Восстановление пропущенных значений экспрессии

Результатом работы алгоритма RMA является построение матрицы уровней экспрессии генов в разных экспериментах. Дальнейшая работа с матрицей нередко осложняется из-за пропусков значений экспрессии.

Пропуски могут быть вызваны повреждением изображения микрочипа, дефектами рабочей поверхности микрочипа, сбоями в функционировании устройств, используемых для процессинга микрочипов, и т.д. Специальный анализ пропущенных значений [33] не позволяет выделить из перечисленных причин доминирующую.

Пропущенные значения экспрессии можно заменить средним уровнем экспрессии исследуемого гена либо нулем (при работе с данными в логарифмическом формате). Более строгие алгоритмы восстановления пропусков учитывают корреляционную структуру экспериментальных данных [33]. В методе k ближайших соседей осуществляется поиск k генов, находящихся на наименьшем расстоянии от i-го гена, значение экспрессии которого в j-м эксперименте пропущено. Недостающее значение принимается равным средневзвешенному уровню экспрессии k ближайших соседей i-го гена в j-м эксперименте. Метод сингулярного разложения основан на построении набора взаимно ортогональных профилей экспрессии («собственных» генов), линейные комбинации которых аппроксимируют профили экспрессии остальных генов [34]. В этом случае профиль экспрессии гена с пропущенным значением раскладывается по наиболее значимым «собственным» генам, и недостающее значение реконструируется с помощью коэффициентов разложения.

Фильтрация экспериментальных данных

Фильтрация данных проводится с целью уменьшения их вариабельности посредством удаления из рассмотрения генов, экспрессия которых измерена с большой погрешностью. Кроме того, из общего объема анализируемых данных исключаются гены, экспрессия которых при варьировании внешних условий или со временем практически не изменяется. Во многих случаях процедура фильтрации уменьшает объем исходных данных в десятки раз, что предъявляет особые требования к этой процедуре. В результирующем множестве по возможности должны остаться только «значимые» гены, а гены, не представляющие интереса в рамках задач исследования, следует отфильтровать.

В отсутствие единых стандартов методы фильтрации обычно выбираются с учетом характера последующего анализа. Простейший алгоритм фильтрации заключается в удалении генов с низкими уровнями экспрессии, которые неотличимы от экспериментальных ошибок. Выбор порогового значения, отделяющего низкий уровень экспрессии от остальных значений, при этом осуществляется произвольно. Проблема выбора порогового значения возникает и при попытке отфильтровать «молчащие» гены, экспрессия которых от эксперимента к эксперименту меняется незначительно [35].

В последние годы получили распространение более строгие методы фильтрации, основанные на использовании критерия χ^2, коэффициентов вариации, ранговых статистик, дисперсионного анализа и др. [29, 36, 37]. К сожалению, основное достоинство подобных алгоритмов — возможность оценки точности фильтрации по табулированным значениям используемых статистик проявляется лишь в случае нормального распределения исходных величин. Более универсальный подход к оцениванию основывается на методах рандомизации [38], которые широко применяются при поиске генов с достоверно изменяющейся экспрессией. Общая идея методов рандомизации состоит в случайном перемешивании исходных экспериментальных значений и сопоставлении количественных характеристик рандомизованного и исходного массивов.

Анализ изменений экспрессии

Основная задача такого анализа заключается в нахождении генов, которые достоверно меняют свою экспрессию со временем либо при изменении условий жизнедеятельности клетки. Простейший способ выявить гены, экспрессия которых различна в двух группах экспериментов, — это сравнить отношение уровней экспрессии с некоторым пороговым значением [18, 39, 40]. К сожалению, этот метод крайне чувствителен к нормировке данных и к сильным различиям дисперсий значений экспрессии в разных экспериментах [41, 42]. Кроме того, остается открытым вопрос о выборе порогового значения. Усовершенствованный алгоритм, также оперирующий отношением уровней экспрессии [24], базируется на линейной модели, которая учитывает различные источники вариабельности экспрессии. Недостатком алгоритма являются значительные погрешности в случае небольших выборок и при высокой неоднородности экспериментального материала [43].

Более последовательно задача поиска генов с достоверно изменяющейся экспрессией может быть сформулирована в терминах проверки статистических гипотез. Нулевая гипотеза H_0 для отдельного гена соответствует тому, что среднее значение его экспрессии в первой группе из p_1 экспериментов не отличается от такового во второй группе из p_2 экспериментов. Для ее проверки могут использоваться традиционные ранговые критерии [44] и методы, предложенные для обработки результатов экспериментов с микрочипами [31, 41, 45], однако наиболее распространенным подходом остается двухвыборочный t-тест.

В своем обычном виде [46] t-тест сводится к построению статистики

$$Z_i = (\hat{y}_{i(1)} - \hat{y}_{i(2)})/(s^2_{i(1)}/p_1 + s^2_{i(2)}/p_2)^{1/2}$$

где $\hat{y}_{i(m)}$ и $s^2_{i(m)}$ среднее значение и оценка дисперсии экспрессии i-го гена в m-й группе экспериментов ($m = 1,2$).

Если погрешности измерения экспрессии генов имеют нормальное распределение, а их дисперсия одинакова для всех генов, статистика Z_i будет подчиняться стандартному t-распределению. Для реальных экспериментальных массивов подобные допущения о вероятностной структуре данных не выполняются даже после перехода в логарифмическую шкалу, которая способствует выравниванию дисперсий и приближает распределение к нормальному [47]. При анализе экспрессий в популяции генов t-тест сохраняет устойчивость к неоднородностям дисперсий, однако обладает недостаточной мощностью и может завышать значения Z_i для низковариабельных генов, изменение экспрессии которых не является достоверным [31, 48].

Вычисление же *t*-статистики по всему массиву даных в предположении гомогенности дисперсий [49] дает смещенные оценки, как только это допущение нарушается.

Существует несколько разновидностей *t*-теста, которые стабилизируют разброс дисперсий экспрессии при переходе от одного гена к другому [31], либо используют средневзвешенные дисперсии экспрессии всех или отдельных генов [50], либо базируются на байесовском подходе, что позволяет учесть изменения экспрессии большого числа генов [51]. Кроме того, для выявления генов с достоверно изменяющейся экспрессией предложены регрессионные алгоритмы [48] и метод смешанных вероятностных моделей [49].

Аппарат теории проверки гипотез (в данном случае гипотезой является допущение о неслучайности изменения экспрессии определенных генов при варьировании экспериментальных условий) позволяет оценить статистическую достоверность выделенного множества генов с изменившейся экспрессией. Обычно ее характеризуют вероятностью ошибки первого рода, называемой *p*-значением, или уровнем значимости. Процедура определения *p*-значения состоит в отнесении к числу генов с изменившейся экспрессией гена, у которого это изменение недостоверно.

Между тем с практической точки зрения гораздо важнее выделить все гены с изменившейся экспрессией, пусть даже ценой ошибочного попадания в их число нескольких лишних генов. Более того, поскольку в большинстве случаев дальнейшему анализу подвергаются именно дифференциально-экспрессированные гены, средняя доля «молчащих» генов, которые могли быть ошибочно идентифицированы как изменившие свою экспрессию, становится мало информативной. Интерес представляет средняя доля ошибок в уже выделенном массиве генов с изменившейся экспрессией, называемая *q*-значением [52]. В [53] предложена процедура FDR (False Discovery Rate), позволяющая вычислить *q*-значения по известным *p*-значениям, которая с успехом может быть применена к анализу экспрессии генов [31, 54]. Несмотря на все еще встречающуюся путаницу между *p*- и *q*-значениями, оперирование *q*-значениями постепенно становится общепринятым.

На практике *p*-значения, используемые при проверке гипотезы H_0, могут быть найдены по распределению вероятностей выбранной статистики, которое моделируется методами рандомизации. Рандомизационные тесты позволяют воспроизвести неизвестную структуру исходных данных при минимальных априорных допущениях. Обычно они сводятся к многократным перестановкам номеров экспериментов и вычислению значений построенной статистики для рандомизированных массивов значений экспрессии [29, 31]. Эффективность этих тестов напрямую зависит от числа экспериментов, которое должно быть достаточным для достижения требуемого уровня значимости. Точность оценивания повышается при наличии нескольких реплик для каждого эксперимента [42]. Несомненное достоинство методов рандомизации — возможность коррекции на множественный характер теста (одновременно анализируется несколько тысяч генов) с учетом сложных корреляционных взаимоотношений между профилями экспрессии разных генов [31, 55].

Обобщением *t*-теста на большее число экспериментальных условий является дисперсионный анализ и *F*-статистика [56, 57]. Простейшая линейная модель дисперсионного анализа для экспрессии *g*-го гена на *r*-м участке *i*-го микрочипа при *k*-м экспериментальном условии и использовании *j*-го красителя-метки может быть записана так

$$y_{ijkgr} = \mu + A_i + D_j + AD_{ij} + z_{ijkgr}$$

где μ — общее среднее значение экспрессии; A_i, D_j и AD_{ij} — не зависящие от конкретного гена эффекты, вносимые микрочипом и меткой по отдельности и совместно. Слагаемое z_{ijkgr} суммирует эффекты, связанные с конкретным геном: $z_{ijkgr} = G_g + AG_{igr} + DG_{jg} + VG_{kg} + \varepsilon_{ijkgr}$ (G_g — среднее значение экспрессии *g*-го гена; AG_{igr} — различие в уровнях экспрессии гена на разных микрочипах; DG_{jg} — специфические для данного гена эффекты метки; VG_{kg} — различие в уровне экспрессии гена, измеряемого в разных экспериментах; ε_{ijkgr} — случайная погрешность).

Доказательство достоверности различий экспрессии гена в нескольких экспериментах эквивалентно проверке гипотезы о том, что все слагаемые VG_{kg} равны нулю, и может быть проведено с использованием *F*-статистики. В работе [42] предложено несколько вариантов построения такой статистики, которые позволяют сопоставить вариабельность экспрессии отдельного гена с общей дисперсией, рассчитанной для всего массива генов. Оценка может быть дана с помощью алгоритмов рандомизации.

Классификация профилей экспрессии

В общей постановке проблема классификации заключается в разбиении анализируемой совокупности объектов на сравнительно небольшое число однородных групп. При анализе результатов микрочиповых экспериментов в роли таких объектов выступают профили экспрессии, отражающие изменение транскрипции в зависимости от фазы клеточного цикла, концентраций веществ или иных внешних условий, либо профили экспрессии, полученные на нескольких клеточных образцах. В соответствии с этим задачи классификации разбиваются на две категории: классификация генов и классификация образцов (экспериментов). Классификация генов заключается в выделении групп генов, имеющих общие механизмы регуляции транскрипции и кодирующие белки одной метаболической цепи [43, 58]. Классификация образцов используется в дифференциальной диагностике заболеваний, при тестировании лекарственных препаратов и в иных в медицинских приложениях [5, 59, 60].

Процедуру классификации генов удобно рассматривать как нахождение компактных групп объектов в признаковом пространстве, размерность которого равна числу экспериментов. В таком пространстве *i*-му гену соответствует точка с координатами X_{i1}, ..., X_{ip}, где X_{ij} — значение его экспрессии в *j*-м эксперименте, *p* — общее число экспериментов. Процедура классификации разбивает анализируемую совокупность точек на группы (кластеры) таким образом, чтобы точки из одного кластера оказались ближе друг к другу, чем

к точкам из других групп. При этом близость точек может быть определена различными способами: помимо привычного евклидова расстояния в алгоритмах кластеризации используются коэффициент корреляции, манхэттенское расстояние, расстояние Махаланобиса, расстояние Чебышева и другие метрики [61]. Выбор метрики имеет существенное значение, поскольку даже при фиксированной процедуре кластеризации разные метрики приводят к разным разбиениям.

Для классификации профилей экспрессии применяются десятки алгоритмов [32, 62, 63], часть из которых была разработана специально для анализа экспрессии генов. Параметрические методы базируются на определенных предположениях о вероятностной структуре исследуемых данных [61], а классификация с обучением — на имеющейся априорной биологической информации о существующих связях между отдельными генами. Поскольку обычно такая информация отсутствует, наибольшее распространение получили непараметрические процедуры [64]. Наиболее популярными среди них являются иерархические алгоритмы.

В иерархической агломеративной кластеризации первоначально каждый объект (профиль экспрессии) отождествляется с отдельным классом. Работа алгоритма начинается с поиска в кластеризуемом массиве двух самых близких точек, которые в результате объединения заменяются на новый класс. На каждом следующем шаге два ближайшие друг к другу класса объединяются в один. Классификация продолжается до тех пор, пока все точки не будут объединены в один кластер. Взаимосвязи между объектами наглядно представляются в виде дерева (дендрограммы), длина ветвей которого отражает близость объектов в относительной шкале. Данный алгоритм был впервые применен к анализу микрочипов в работе [65], после чего для классификации профилей экспрессии стал использоваться повсеместно.

Несомненными преимуществами метода иерархической кластеризации являются простота реализации и возможность визуализации результатов. В то же время алгомеративные процедуры нередко дают ошибочные результаты из-за сильной зависимости первых шагов алгоритма от локальных сгущений, которые могут присутствовать в исходном экспериментальном массиве: отсутствие итерационных шагов приводит к накапливанию небольших ошибок, возникших в самом начале процедуры классификации [63]. Кроме того, с увеличением размеров кластеров их центры масс могут сильно отличаться от реальных классифицируемых объектов, в результате чего картина объединения генов становится менее достоверной [27]. Наконец, иерархические алгоритмы имеют тенденцию к неоправданному дроблению больших кластеров [32] и наиболее эффективны для задачи разбиения данных только по одному признаку [66].

В отличие от иерархической кластеризации, так называемые алгоритмы разбиения — метод k-средних, самоорганизующиеся карты (SOM), разбиение по многомерным медианам (PAM) и др. — разбивают данные на заданное число классов методом последовательных приближений, каждый раз формируя разбиение заново [61, 64]. Например, в методе k-средних [62] предполагается, что число классов k известно заранее и путем итерационного перемещения объектов между классами строится разбиение, минимизирующее сумму расстояний от объектов до центров тех классов, к которым они принадлежат. Продуктивность алгоритма k-средних в анализе экспрессии генов была впервые продемонстрирована при изучении изменения экспрессии генов в ходе клеточного цикла дрожжей *S. cerevisiae* [67], после чего он также получил широкое распространение.

Помимо очевидности основной идеи и простоты реализации преимуществом метода k-средних является быстрая сходимость результатов, которая нередко достигается за 5—10 итераций [61]. Проблемы, возникающие при его практическом применении, связаны с необходимостью предварительного задания числа классов, которое в большинстве случаев неизвестно, с неопределенностью в расположении начальных центров классов, которое влияет на достоверность результатов и их устойчивость к малым возмущениям исходного экспериментального массива, а также со склонностью алгоритма выделять шарообразные кластеры одинакового размера, далеко отстоящие друг от друга [61, 63]. Теми же недостатками обладает похожий на процедуру k-средних метод SOM [68], который в дополнение к величине k (число классов) требует задания других управляющих параметров. Его частое использование для анализа экспрессии генов [66] можно объяснить только удобством визуализации сложных многомерных профилей экспрессии в двух- или трехмерном пространстве [32].

Принципиальной альтернативой эвристическим методам являются процедуры, основанные на предположении о том, что данные представляют собой выборку из смеси вероятностных распределений, каждое из которых соответствует отдельному кластеру [69, 70]. К их недостаткам следует отнести использование априорных допущений об определенной вероятностной структуре профилей экспрессии генов и крайне низкую сходимость ряда алгоритмов [71].

В дополнение к описанным методам кластеризации для анализа профилей генной экспрессии применяется множество других алгоритмов: процедуры, основанные на математической теории графов [72, 73], методы нечеткой кластеризации [74, 75], кластеризации траекторий [37], двойной кластеризации [76], сжатых центроидов [77], взаимной информации [78], суперпарамагнитной кластеризации [76] и др. Каждый метод имеет свои сильные и слабые стороны, поэтому универсальные рекомендации здесь вряд ли возможны. По-видимому, наилучшей стратегией является применение нескольких алгоритмов к одному и тому же массиву данных и сравнение полученных результатов на основе некоторого критерия. При этом полезной может оказаться визуализация разброса исходных данных в пространстве первых главных компонентов, а также сравнение средних профилей отдельных кластеров с профилями, построенными для генов с известной биологической функцией [79].

Определение числа классов

Помимо выбора алгоритма кластеризация профилей экспрессии требует корректного определения числа классов и оценки качества полученного разбиения.

Проблема априорного выбора числа классов исследуется уже давно [75, 80]. Новый всплеск интереса к

ней обусловлен широким применением технологии микрочипов, поскольку от числа классов в разбиении профилей экспрессии зависит последующая биологическая интерпретация результатов кластеризации, установление функциональных связей между генами, оказавшимися в одном классе, поиск общих механизмов регуляции их транскрипции и т.д.

Первые алгоритмы определения числа классов возникли в рамках методов кластеризации на базе вероятностных моделей [69, 80]. Число классов в этих методах определяется автоматически, однако сфера их применимости ограничена из-за использования допущений об определенных вероятностных свойствах исходных данных. Более того, для многомерных выборок небольшого размера точность вероятностных методов остается невысокой [81], а именно такая ситуация имеет место при анализе результатов микрочиповых экспериментов, где число генов на 1—2 порядка превосходит число экспериментов. В ряде случаев методы, не использующие каких-либо допущений о вероятностной структуре исходных данных [69, 82], неявно все же опираются на определенную априорную модель, например, предполагают высокую компактность классов [83].

Более продуктивными представляются эвристические процедуры, в которых используются показатели компактности классов, вычисляемые на основании полученного разбиения [82, 84, 85]. Алгоритм кластеризации применяется к исходным данным, а предполагаемое число классов k варьируется в некотором диапазоне. Истинное число классов определяется по выполнению для показателя компактности определенного условия [66, 82, 86—88]. Основные недостатки методов, базирующихся на применении показателей компактности, связаны с их многообразием и эвристической природой. Оценки числа классов в одних и тех же экспериментальных массивах значений генной экспрессии, полученные с применением разных показателей компактности, дают разные результаты [82, 85].

Эту трудность частично удается обойти при помощи перевыборочных алгоритмов [82, 85, 88—90]. Число классов может быть определено посредством сравнения внутриклассовой дисперсии разбиения реальных данных с аналогичной величиной для эталонного нулевого разбиения [82] либо путем многократного разбиения массива на обучающую и тестовую подвыборки [83, 89].

Общим недостатком большинства методов определения числа классов является игнорирование статистических свойств величин, характеризующих компактность или иные свойства получаемого разбиения. Между тем непараметрическое оценивание таких свойств необходимо для проверки статистической гипотезы о том, что «истинное» число классов равно выбранному значению. Определенный шаг в этом направлении сделан в работе [89], авторы которой сравнивают разбиение на k кластеров исходного массива и большого числа его подмножеств.

Оценка качества разбиения профилей экспрессии на классы

Качество разбиения отражает вторую часть проблемы, связанной с проверкой способности алгоритма классификации выявить истинную структуру в массиве профилей генной экспрессии: после определения числа классов и кластеризации профилей необходимо убедиться в том, что построенное разбиение в определенном смысле оптимально. Актуальность этой задачи обусловлена тем обстоятельством, что даже для фиксированного числа классов разные алгоритмы кластеризации приводят к разным разбиениям одного и того же экспериментального массива данных [79], а это самым непосредственным образом влияет на последующую биологическую интерпретацию результатов кластеризации.

В течение длительного времени качество классификации определялось по стабильности разбиения в целом, устойчивость же отдельных кластеров стала рассматриваться сравнительно недавно [90—92]. В обоих случаях речь идет о стабильности по отношению к изменению экспериментальных условий либо значений экспрессии, которое моделируется путем рандомизации данных [90, 92—96] или добавления к ним случайной величины [97]. Устойчивость разбиения и отдельных классов можно характеризовать, например, числом пар, которые остались неразделенными после кластеризации возмущенного массива данных. В роли возмущающего фактора может выступать гауссовский шум, моделирующий распределение ошибок в экспериментах с микрочипами [97].

Идея сравнения состава отдельных кластеров в исходном и производных массивах, но уже сгенерированных методом перестановки, получила развитие в работе [90]. Авторы предложили способ выделения наиболее стабильных кластеров, а также наиболее репрезентативных профилей экспрессии для каждого кластера. Аналогичный подход может применяться для определения числа классов в данных.

В исследованиях с применением микрочипов число экспериментов всегда гораздо меньше числа изучаемых генов. Этот факт порой трактуется как явная избыточность экспериментальных данных, что позволяет использовать в дальнейшем анализе только их часть. Подобная точка зрения легла в основу метода оценки стабильности отдельных кластеров [92] путем отбрасывания результатов случайно выбранной части микрочиповых экспериментов. Многократное повторение процедуры и сравнение результатов кластеризации исходных и редуцированных профилей экспрессии позволяет подсчитать долю итераций, в которых данный кластер остался неизменным. В отличие от процедуры, описанной в [97], этот метод не требует подмешивания к данным случайного шума.

В связи с широким распространением иерархических алгоритмов в последнее время появились методы оценки устойчивости кластеров, основанные на отслеживании их перемещения по дендрограмме [91] либо на определении уровня значимости ветвей дендрограммы и отдельных генов в рамках гауссовой модели [93, 95]. Главный недостаток подобных процедур заключается в наличии определенных допущений о вероятностной структуре данных либо в использовании только части исходной выборки [92].

Нерешенные проблемы

Практически на каждом этапе анализа профилей генной экспрессии, измеряемых при помощи микро-

чипов, исследователь сталкивается с проблемой выбора адекватных количественных процедур. Сложность этого выбора обусловлена тем, что с появлением технологии микрочипов биология впервые столкнулась с математическими задачами подобного масштаба [13]. Их решение заставило адаптировать существующие методы обработки данных к анализу профилей экспрессии генов и одновременно стимулировало создание многочисленных новых алгоритмов. Наиболее перспективными среди них являются процедуры, основанные на строгом статистическом оценивании и свободные от априорных допущений о вероятностной структуре экспериментальных данных. Ситуация усугубляется отсутствием в данной области стандартов, появление которых невозможно без скрупулезного тестирования алгоритмов на разнообразном модельном и экспериментальном материале. К сожалению, эта работа еще далека от своего завершения.

Самостоятельную проблему представляет моделирование механизмов регуляции транскрипции и ответа генетического аппарата клетки на изменение внешних условий. Результаты моделирования позволили бы воспроизвести реальные массивы профилей экспрессии и построить адекватную вероятностную модель экспериментальных данных, которая впоследствии могла бы стать основой параметрических методов анализа.

Не стоит сбрасывать со счетов и ограниченность самой технологии исследования с использованием микрочипов: некоторые регуляторные процессы осуществляются на посттранскрипционном уровне, а значит, остаются за рамками микрочиповых экспериментов. Это обстоятельство заставляет относиться к результатам анализа профилей экспрессии и, главное, к их последующей интерпретации с известной осторожностью.

В сложившихся условиях наибольшие перспективы можно ожидать от подходов, основанных на сопоставлении результатов микрочиповых экспериментов с биологическими и клиническими данными, поиске регуляторных участков генов, изучении клеточных сигнальных путей и метаболических сетей. Решение проблем количественной обработки данных и сопоставление результатов микрочиповых экспериментов с существующими сведениями из области биохимии, биологии клетки, молекулярной генетики и структурной биологии будут способствовать реализации потенциала, скрытого в новых экспериментальных технологиях исследования генома.

ЛИТЕРАТУРА

1. *Elvidge G.* Pharmacogenomics, 2006, v. 7, p. 123—134.
2. *Golub T.R., Slonim D.K., Tamayo P. e. a.* Science, 1999, v. 286, p. 531—537.
3. *AlonU., Barkai N., Notterman D.A. e. a.* Proc. Natl. Acad. Sci. USA, 1999, v. 96, p. 6745—6750.
4. *Shioda T.* J. Environ. Pathol. Toxicol. Oncol., 2004, v. 23, p. 13—31.
5. *Debouck C., Goodfellow P.N.* Nat. Genet. Suppl., 1999, v. 21, p. 48—50.
6. *Robson P.* Trends Biotechnol., 2004, v. 22, p. 609—612.
7. *Pappas C.T., Sram J., Moskvin O.V. e. a.* J. Bacteriol., 2004, v. 186, p. 4748—4758.
8. *Hacia J.* Nat. Genet. Suppl., 1999, v. 21, p. 42—47.
9. *Saiki R.K., Bugawan T.L., Horn G.T. e. a.* Nature, 1986, v. 324, p. 163—166.
10. *Alwine J.C., Kemp D.J., Stark G.R.* Proc. Natl. Acad. Sci. USA, 1977, v. 74, p. 5350—5354.
11. *Velculescu V.E., Zhang L., Vogelstein B. e. a.* Science, 1995, v. 270, p. 484—487.
12. *Schena M., Shalon D., Davis R.W e. a.* Ibid., 1995, v. 270, p. 467—470.
13. *Duggan D.J., Bittner M., Chen Y. e. a.* Nat. Genet. Suppl., 1999, v. 21, p. 10—14.
14. *Hughes T.R., Mao M., Jones A.R. e. a.* Nat. Biotechnol., 2001, v. 19, p. 342—347.
15. *Lipshutz R.J., Fodor S.P., Gingeras T.R. e. a.* Nat. Genet., 1999, v. 21, p. 20—24.
16. *Blanchard A.P., Kaiser R.J., Hood L.E.* Biosensors and Bioelectronics, 1996, v. 11, p. 687—690.
17. *Maskos U., Southern E.M.* Nucl. Acids Res., 1993, v. 21, p. 2267—2268.
18. *Schena M., Shalon D., Heller R. e. a.* Proc. Natl. Acad. Sci. USA, 1996, v. 93, p. 10614—10619.
19. *Evertsz E.M., Au-Young J., Ruvolo M.V. e. a.* Biotechniques, 2001, v. 31, p. 1182—1186.
20. *Murphy D.* Adv. Physiol. Educ., 2002, v. 26, p. 256—270.
21. *Schadt E., Li C., Su C. e. a.* J. Cell Biochem., 2000, v. 80, p. 192—202.
22. *Sebastiani P., Gussoni E., Kohane I.S. e. a.* Statist Sci., 2003, v. 18, p. 33—70.
23. *Yang Y.H., Dudoit S., Luu P. e. a.* Nucl. Acids Res., 2002, v. 30, p. e15.
24. *Li C., Wong W.* Proc. Natl. Acad. Sci. USA, 2001, v. 98, p. 31—36.
25. *Irizarry R.A., Hobbs B., Colin F. e. a.* Biostatistics, 2003, v. 4, p. 249—264.
26. *Irizarry R.A., Bolstad B.M., Colin F. e. a.* Nucl. Acids Res., 2003, v. 31, p. e15.
27. *Quackenbush J.* Nat. Rev. Genet., 2001, v. 2, p. 418—427.
28. *Rhodius V., Van Dyk T.K., Gross C. e. a.* Annu. Rev. Microbiol., 2002, v. 56, p. 599—624.
29. *Dudoit S., Yang Y.H., Callow M.J. e. a.* Statistica Sinica, 2002, v. 12, p. 111—139.
30. *Tseng G.C., Oh M.-K., Rohlin L. e. a.* Nucl. Acids Res., 2001, v. 29, p. 2549—2557.
31. *Tusher V.G., Tibshirani R., Chu G.* Proc. Natl. Acad. Sci. USA, 2001, v. 98, p. 5116—5121.
32. *Jiang D., Tang C., Zhang A.* IEEE Trans. Knowl. Data Eng., 2004, v. 16, p.1370—1386.
33. *Troyanskaya O., Cantor M., Sherlock G. e. a.* Bioinformatics, 2001, v. 17, p. 520—525.
34. *Kuruvilla F.G., Park P.J., Schreiber S.L.* Genome Biol., 2002, v. 3, p. res. 0011.
35. *Fambrough D., McClure K., Kazlauskas A. e. a.* Cell, 1999, v. 97, p. 727—741.
36. *Mills J.C., Gordon J.I.* Nucl. Acids Res., 2001, v. 29, p. e72.
37. *Phang T.L., Neville M.C., Rudolph M. e. a.* Pac. Symp. Biocomput., 2003, p. 351—362.
38. *Эфрон Б.* Нетрадиционные методы многомерного статистического анализа. М.: Финансы и статистика, 1988.
39. *DeRisi J.L., Iyer V.R., Brown P.O.* Science, 1997, v. 278, p. 680—686.
40. *Sabatti C., Karsten S.L., Geschwind D.H.* Math. Biosci., 2002, v. 176, p. 17—34.
41. *Ideker T., Thorsson V., Siegel A.F. e. a.* J. Comput. Biol., 2000, v. 7, p. 805—817.

42. *Cui X., Churchill G.A.* Genome Biol., 2003, v. 4, p. 210.
43. *Slonim D.K.* Nat. Genet., 2002, v. 32, p. 502—508.
44. *Рунион Р.* Справочник по непараметрической статистике. М.: Финансы и статистика, 1982.
45. *Pan W., Lin J., Le C.* Funct. Integr. Genom., 2003, v. 3, p. 117—124.
46. *Pan W.* Bioinformatics, 2002, v. 18, p. 546—554.
47. *Quackenbush J.* Nat. Genet., 2002, v. 32, Suppl., p. 496—501.
48. *Thomas J.G., Olson J.M., Tapscott S.J. e. a.* Genome Res., 2001, v. 11, p. 1227—1236.
49. *Tanaka T.S., Jaradat S.A., Lim M.K. e. a.* Proc. Natl. Acad. Sci. USA, 2000, v. 97, p. 9127—9132.
50. *Baldi P., Long A.D.* Bioinformatics, 2001, v. 17, p. 509—519.
51. *Lönnstedt I., Speed T.P.* Statistica Sinica, 2002, v. 12, p. 31.
52. *Storey J.D.* J. Royal Stat. Soc. Ser. B, 2002, v. 64, p. 479—498.
53. *Benjamini Y., Hochberg Y.* Ibid., 1995, v. 57, p. 289—300.
54. *Storey J., Tibshirani R.* Proc. Natl. Acad. Sci. USA, 2003, v. 100, p. 9440—9445.
55. *Westfall P.H., Young S.S.* Resampling-Based Multiple Testing. New York: John Wiley & Sons, 1993.
56. *Kerr M.K., Martin M., Churchill G.A.* J. Comput. Biol., 2000, v. 7, p. 819—837.
57. *Pritchard C.C., Hsu L., Delrow J. e. a.* Proc. Natl. Acad. Sci. USA, 2001, v. 98, p. 13266—13271.
58. *DeJong H.* J. Comput. Biol., 2002, v. 9, p. 67—103.
59. *Yeoh E.,Ross M.E., Shurtleff S.A. e. a.* Cancer Cell, 2002, v. 1, p.133—143.
60. *Stegmaier K., Ross K.N., Colavito S.A. e. a.* Nat. Genet., 2004, v. 36, p. 257—263.
61. *Jain A.K., Dubes R.C.* Algorithms for clustering data. Englewood Cliffs, NJ: Prentice-Hall, 1988.
62. *Shamir R., Sharan R.* In: Current Topics in Computational Biology. Eds. Jiang T., Smith T., Xu Y., Zhang M. Boston: MIT Press, 2002, p. 269—299.
63. *Valafar F.* Ann. N.Y. Acad. Sci., 2002, v. 980, p. 41—64.
64. *Kaufman L., Rousseeuw P.J.* Fitting Groups in Data. An Introduction to Cluster Analysis. New York: Wiley, 1990.
65. *Eisen M.B., Spellman P.T., Brown P.O. e. a.* Proc. Natl. Acad. Sci. USA, 1998, v. 95, p. 14863—14868.
66. *Tamayo P., Slonim D., Mesirov J. e. a.* Ibid., 1999, v. 96, p. 2907—2912.
67. *Tavazoie S., Hughes J.D., Cambell M.J. e. a.* Nat. Genet., 1999, v. 22, p. 281—283.
68. *Kohonen T.* Self-Organizing Maps. Berlin: Springer, 1997.
69. *Yeung K.Y., Fraley C., Murua A. e.a.* Bioinformatics, 2001, v. 17, p. 977—987.
70. *Ghosh D.* Ibid., 2001, v. 17, p. 275—286.
71. *Fraley C., Raftery A.E.* The Computer J., 1998, v. 41, p. 578—588.
72. *Hartuv E., Schmitt A., Lange J. e. a.* In RECOMB99: Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB'99), Lyon, France, 1999, p. 188—197.
73. *Sharan R., Maron-Katz A., Shamir R.* Bioinformatics, 2003, v. 19, p. 1787—1799.
74. *Dunn J.C.* J. Cybern., 1974, v. 4, p. 95—104.
75. *Halkidi M., Batistakis Y., Vazirgiannis M.* Intel. Inform. Syst., 2001, v. 17, p. 107—145.
76. *Getz G., Levine E., Domany E.* Proc. Natl. Acad. Sci. USA, 2000, v. 97, p. 12079—12084.
77. *Tibshirani R., Hastie T., Narasinmhan B. e. a.* Ibid., 2002, v. 99, p. 6567—6572.
78. *Butte A.J., Hohane I.S.* Pac. Symp. Biocomp., 2000, p. 418—429.
79. *Datta S. e.a.* Bioinformatics, 2003, v. 19, p. 459—466.
80. *Banfield J., Raftery A.E.* Biometrics, 1993, v. 49, p. 803—821.
81. *Kass R.E., Raftery A.E.* J. Am. Stat. Assoc., 1995, v. 90, p. 773—795.
82. *Tibshirani R., Walther G., Hastie T.* J. Royal Stat. Soc. B, 2001, v. 63, p. 411—423.
83. *Lange T., Roth V., Braun M.L. e. a.* Neural Comput., 2004, v. 16, p. 1299—1323.
84. *Milligan G.W., Cooper M.C.* Psychometrika, 1985, v. 50, p. 159—179.
85. *Dudoit S., Fridlyand J.* Genome Biol., 2002, v. 3, p. res. 0036.1—0036.21.
86. *Calinski R., Harabasz J.* Commun. Stat., 1974, v. 3, p. 1—27.
87. *Krzanowski W., Lai Y.* Biometrics, 1985, v. 44, p. 23—34.
88. *Giurcăneanu C.D., Tăbuş I.* Eur. J. Appl. Signal Proc., 2004, v. 1, p. 64—80.
89. *Ben-Hur A., Elisseeff A., Guyon I.* Pac. Symp. Biocomput., 2002, p. 6—17.
90. *Monti S., Tamayo P., Mesirov J. e. a.* Machine Learning, 2003, v. 52, p. 91—118.
91. *Bittner M., Meltzer P., Khan J. e. a.* Nature, 2000, v. 406, p. 536—540.
92. *Smolkin M., Ghosh D.* BMC Bioinformatics, 2003, v. 4, p. 36.
93. *Zhang K., Zhao H.* Funct. Integr. Genomics, 2000, v. 1, p. 156—173.
94. *Bhattacharjee A., Richards W.G., Staunton J. e. a.* Proc. Natl. Acad. Sci. USA, 2001, v. 98, p. 13790—13795.
95. *Kerr M.K., Churchill G.A.* Ibid., 2001, v. 98, p. 8961—8965.
96. *Garge N.R., Page G.P., Sprague A.P. e. a.* BMC Bioinformatics, 2005, v. 6, p. S10.
97. *McShane L.M, Radmacher M.D, Friedlin B. e. a.* Bioinformatics, 2002, v. 18, p. 1462—1469.