БИОИНФОРМАТИКА

УДК 004.048

Н.А. Новоселова, И.Э. Том

МЕТОД ПОСТРОЕНИЯ КЛАСТЕРОВ ГЕНЕТИЧЕСКИХ ДАННЫХ

Предлагается метод построения кластеров (функциональных модулей) генетических данных, основанный на использовании рандомизированных матриц. Для построения кластеров выполняется выделение и анализ главных компонент матрицы корреляций генных профилей. В качестве конечных выбираются главные компоненты, которые соответствуют собственным значениям, значимо отличающимся от полученных при анализе случайным образом сгенерированной корреляционной матрицы (рандомизированной). В сравнительном вычислительном эксперименте с аналогами метод показал свое преимущество в возможности выделять статистически значимые кластеры малых и больших размеров, способности отфильтровывать неинформативные признаки, а также получать биологически интерпретируемые функциональные модули, адекватные реальной структуре данных.

Введение

Новые технологии секвенирования генных последовательностей, которые позволяют получить большое количество информации о геноме различных организмов, способствовали развитию различных компьютерных методов, предназначенных для поиска схожих последовательностей ДНК, предсказания структуры белков и функциональных свойств генов. Особенно бурно данное направление стало развиваться после появления таких геномных технологий, как олигонуклеотидные и кДНК-микрочипы (кДНК – комплементарная ДНК). Типичный набор данных экспрессии генов, полученный с использованием геномной технологии, включает, как правило, сотни образцов для тысяч и десятков тысяч генов, которые содержат большой объем скрытой информации. Для ее обнаружения необходимо использовать специальные методы, которые в дополнение к анализу ДНК-последовательностей, заключающемуся в поиске генов и их положения в последовательности, позволяют выделить функциональные модули схожих генных профилей. Гены, относящиеся к одному и тому же кластеру, обычно отвечают за определенный физиологический процесс или относятся к одному и тому же молекулярному комплексу.

1. Обзор существующих методов

Для выделения функциональных модулей в данных генной экспрессии были предложены различные методы машинного обучения, которые позволяют строить клеточные сети. Наиболее широко используемыми являются методы построения булевых сетей, байесовских сетей, иерархический кластерный анализ, кластерный метод *k*-средних, самоорганизующиеся карты Кохонена (SOM) и ассоциативные корреляционные сети.

Метод построения булевых сетей [1] использует упрощенное представление генных связей, где 1 кодирует наличие связи, а 0 – ее отсутствие. Байесовские сети позволяют графически представить зависимость между генами на основе оценки условных вероятностей. При иерархической кластеризации [2] выполняется итерационный процесс группировки генов, когда первоначально группируются гены с большим значением корреляции. Однако этот алгоритм может сходиться к локальному минимуму, так как объединение генов в кластеры происходит в прямом направлении и решение, принятое на некоторой итерации алгоритма, не может быть изменено на последующих итерациях. SOM [3] представляет собой нейросетевой кластерный алгоритм, для которого требуется задание начального числа кластеров, что почти всегда затруднительно. Ассоциативные корреляционные сети широко используются для идентификации клеточных сетей, поскольку обладают вычислительной эффективностью и способностью

справляться с особенностями данных микрочипов (зашумленностью, высокой размерностью и значительным количеством пропусков). Недостатком таких сетей является произвольным образом задаваемые пороговые значения для установления связей в сети, что приводит к субъективности в определении структуры и топологии сетей. Известны различные программные комплексы кластеризации и визуализации данных микрочипов, например: affy, cclust, cluster, mcluster, hybridHclust, SOM-пакет в среде R [4], Bioconductor [5] и Cluster3.0/Tree view [6], вебсистемы, такие как cyberT [7], SNOMAD [8]. Недостатки описанных выше методов не позволяют выявлять наиболее значимые кластеры и отфильтровывать шумы в данных, поэтому задача построения транскрипторных сетей и обнаружения несмещенных оценок биологических связей в данных генной экспрессии продолжает оставаться актуальной и исследованиям в этом направлении уделяется большое внимание.

Последующие разделы посвящены подробному описанию предлагаемого метода, итерационной процедуры процесса выделения функциональных модулей, процедуры оценки стабильности полученных кластеров, а также результатам тестирования метода на наборах искусственно сгенерированных и реальных данных.

2. Описание метода построения кластеров генетических данных

Определим значение экспрессии гена i, $i = 1, \ldots, N$, в нескольких экспериментах как

$$W_i(s) = \ln\left(\frac{Es_i(s)}{Ec_i(s)}\right),$$

где $Es_i(s)$ – значение экспрессии гена i в эксперименте s, $s = 1, ..., K$; $Ec_i(s)$ – контрольное значение. Для того чтобы учесть различные уровни экспрессии для разных генов, проводится стандартизация генных профилей:

$$w_i(s) = \frac{W_i(s) - \overline{W}_i}{\sigma_i},$$

где $\sigma_i = \sqrt{\overline{W_i^2} - (\overline{W}_i)^2}$ – стандартное отклонение переменной W_i, а $\overline{W}_i$ – среднее значение по различным экспериментам для гена i. Используя стандартизированную матрицу M размерности $N \times K$, рассчитаем матрицу кросс-корреляции C:

$$C = \left(\frac{1}{K}\right) MM^T.$$

Коэффициент корреляции Пирсона C_{xy} между генами x и y, заданными K-мерным вектором, может быть также рассчитан с использованием формулы

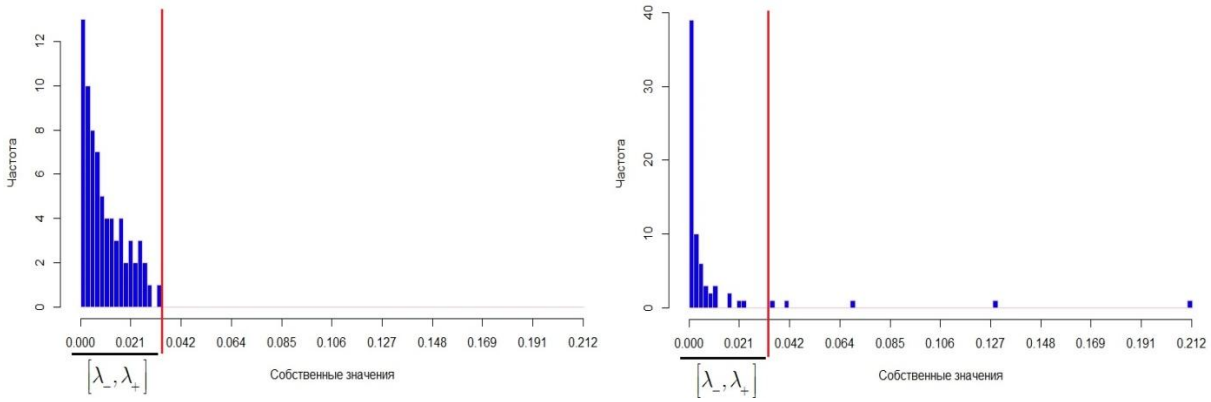
$$C_{xy} = \frac{\sum\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\left(K-1\right)s_x s_y},$$

где s_x, s_y – стандартные отклонения.

Значение коэффициента корреляции изменяется в диапазоне от –1 до +1. При $C_{ij} = 0$ корреляция между генами отсутствует. Однако оценить точное значение коэффициента корреляции генных профилей затруднительно в связи с особенностями данных микрочипов. Кросс-корреляция между любой парой генов не всегда является константой, а может изменяться во времени или при различных условиях экспериментов. Кроме того, в условиях ограниченного количества экспериментов на значение корреляции оказывают влияние ошибки эксперимента,

или экспериментальный шум. Чтобы отфильтровать случайную компоненту при оценке значений матрицы кросс-корреляции, проводится сравнительный анализ собственных значений данной матрицы и значений матрицы, случайным образом сгенерированной (рандомизированной). Статистические показатели, соответствующие рандомизированной матрице, не принимаются в дальнейшее рассмотрение и помечаются как шум. Собственные значения исходной матрицы, отличающиеся от таковых для рандомизированной матрицы, рассматриваются в качестве отправных для дальнейшего анализа и построения транскрипторной сети.

Для того чтобы определить значимые для дальнейшего анализа собственные значения кросс-корреляционной матрицы, анализируются вероятностные распределения $P^C(\lambda)$ и $P^R(\lambda)$ собственных значений λ_i соответственно исходной матрицы кросс-корреляции C и случайным образом сгенерированной матрицы R. Собственные значения сортируются по возрастанию, т. е. $\lambda_i < \lambda_{i+1}$. На рис. 1 показаны распределения $P^C(\lambda)$ и $P^R(\lambda)$ для набора данных по анализу клеточного цикла дрожжевых грибков [9]. Из рисунка следует, что множество собственных значений матрицы C находится в диапазоне значений $[\lambda_-, \lambda_+]$, рассчитанных для матрицы R, но с несколькими значениями, выходящими за пределы нижней λ_- и верхней λ_+ границ. Эти собственные значения и соответствуют реальной информации о корреляции между признаками.



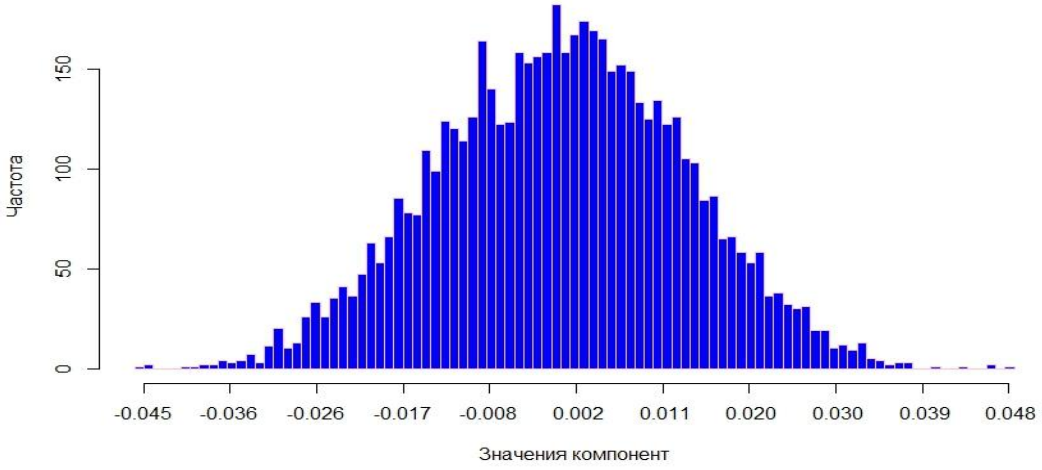
а) *б)*

Рис. 1. Гистограммы плотностей распределения собственных значений:
а) рандомизированная матрица; *б)* реальная корреляционная матрица

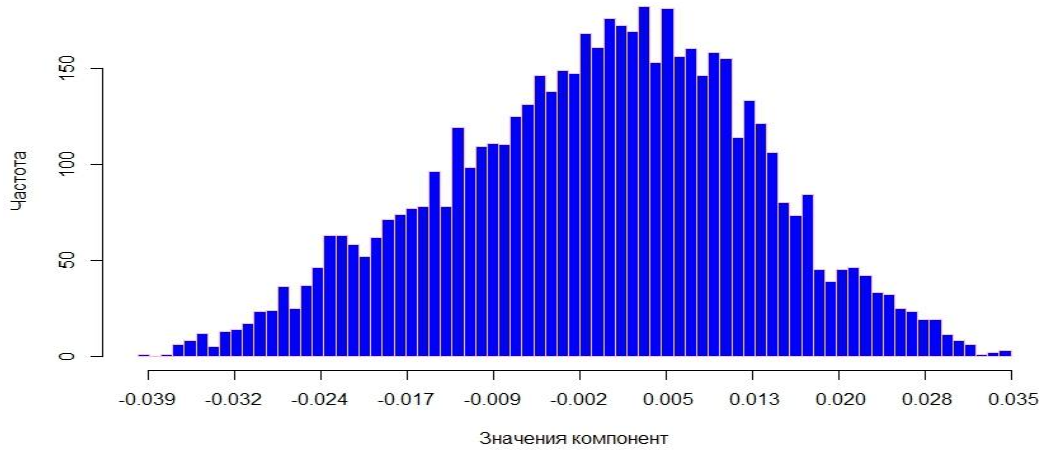
Таким образом, используя собственные значения, можно отделить реальную корреляцию от случайной и выполнить процесс фильтрации шума, что особенно актуально при анализе данных, полученных с помощью технологии микрочипов. Эти данные, как правило, содержат малое число экспериментов по отношению к количеству генов, для которых выполняется измерение экспрессии. Экспериментально подтверждено, что при увеличении количества экспериментов или временного ряда собственные значения корреляционной матрицы все больше отклоняются от диапазона соответствующих значений, полученных для рандомизированной матрицы. Предполагается, что при увеличении числа экспериментов K удастся выявить наиболее трудноразличимые группы генов, которые невозможно определить с помощью стандартных кластерных методов. Однако на практике выполнение большого количества экспериментов является затратным по времени и требует привлечения значительных материальных ресурсов. Для построения функциональных модулей генов можно использовать только собственные значения, которые выходят за диапазон собственных значений рандомизированной матрицы. Эти функциональные модули и представляют собой кластеры генов, которые выделены из зашумленных данных. Величина дисперсии, привносимая каждым собственным вектором, определяется как отношение соответствующего собственного значения к сумме всех собственных значений корреляционной матрицы. Как и при анализе глав-

ных компонент, где главные факторы объясняют большую часть вариации в данных, в предложенном методе только бóльшие собственные значения и соответствующие собственные векторы рассматриваются для анализа групп генов. Остальные собственные значения являются неинформативными.

Рассмотрим распределение компонент собственных векторов, соответствующих информативным и неинформативным собственным значениям. Пусть имеется N собственных векторов u^i, $i = 1, \ldots, N$. Каждый собственный вектор u_i имеет N компонент, соответствующих генам анализируемой матрицы. Вероятностное распределение компонент собственного вектора для собственных значений реальной корреляционной матрицы и рандомизированной матрицы, заданной гауссовым распределением с нулевым средним значением и единичной дисперсией, показано на рис. 2.



а)



б)

Рис. 2. Гистограммы распределения компонент собственного вектора, соответствующего:
а) собственному значению в диапазоне значений рандомизированной матрицы;
б) выходящему из диапазона $[\lambda_-, \lambda_+]$ собственному значению реальной матрицы

Вероятностное распределение компонент собственного вектора λ_k из диапазона $\lambda_- \le \lambda_k \le \lambda_+$ собственных значений рандомизированной матрицы для данных по анализу клеточного цикла дрожжевых грибков хорошо согласуется с нормальным распределением (рис. 2, *а*).

Распределение компонент собственного вектора, соответствующего собственному значению, выходящему из диапазона, отличается от распределения Гаусса (рис. 2, *б*). Вычислительные эксперименты показали, что при приближении собственного значения к диапазону значений рандомизированной матрицы их распределение стремится к нормальному.

После того как определены собственные векторы, которые соответствуют информативным собственным значениям (выходящим из диапазона значений для рандомизированной матрицы), компоненты векторов преобразуются в весовые коэффициенты путем перемножения их значений на квадратный корень из соответствующих собственных значений. Каждый собственный вектор далее представляет собой один кластер генов. Большее значение весового коэффициента означает большее влияние соответствующего гена на формирование собственного вектора (другими словами, доминирующую позицию соответствующего гена в кластере). Чтобы упростить структуру собственного вектора и облегчить интерпретацию кластера генов, можно применить процедуру ортогонального вращения собственных векторов. Для выполнения вращения используется метод VARIMAX [10], который позволяет преобразовать основные координатные оси таким образом, чтобы каждый собственный вектор содержал меньшее количество больших значений весовых коэффициентов и большее количество нулей или малых значений весовых коэффициентов. С точки зрения биологии это означает, что каждый ген является составной частью кластеров или оказывает влияние только на малое их количество и каждый кластер состоит из меньшего количества доминирующих генов, чем до вращения. Матрица вращения R в общем виде определяется как

$$R = \begin{bmatrix} \cos\theta_{i,i} & \cos\theta_{i,j} \\ \cos\theta_{j,i} & \cos\theta_{j,j} \end{bmatrix},$$

где $\theta_{i,j}$ – угол вращения, преобразующий старые оси в новые.

В предлагаемом методе реализована итерационная процедура выделения функциональных модулей генетических данных. Согласно процедуре функциональные модули выделяются сначала из всего анализируемого набора данных, а затем из каждого выделенного генного кластера. Результатом такой процедуры является иерархическая кластерная структура, подобная структуре, получаемой стандартными методами иерархической кластеризации. Отличием является то, что на каждом последующем шаге процедуры в рассмотрение принимается не все содержимое кластера, а только его часть, соответствующая наиболее значимым собственным значениям корреляционной матрицы.

3. Оценка стабильности кластеров

Оценка стабильности выделенных кластеров основана на оценке собственных значений корреляционной матрицы. Новые значения координат экспериментов s рассчитываются с использованием компонент собственных векторов u^i, где каждая из координат $z^i(s)$ представляет собой значение эксперимента в пространстве собственных векторов:

$$z^i(s) = \sum_{k=1}^{N} u_k^i W_k(s).$$

Стабильность i-го кластера генов может быть оценена через дисперсию всех экспериментов z^i по координате собственного вектора i. Дисперсия напрямую ассоциируется с соответствующим собственным значением:

$$Stab(u^i) = \sigma^2(z^i) = (u_i)^T C u^i = \lambda_i,$$

где $i = 1, \ldots, N$.

Таким образом, генный кластер, порожденный собственным вектором с бóльшим собственным значением, является менее стабильным, значение дисперсии соответствует согласованности генов в кластере по всем экспериментам и предоставляет дополнительную информацию о качестве кластера и его свойствах.

4. Результаты тестирования метода на искусственно сгенерированном наборе данных

Для тестирования предлагаемого метода построения кластеров генетических данных, основанного на теории рандомизированных матриц, и проведения его сравнительного анализа с известными методами разработана методика генерации случайных наборов данных, которые имитируют данные экспрессии генов. В соответствии с методикой данные генерируются по следующим правилам: 1) согласно нормальному распределению генерируются несколько многомерных кластеров генов, причем предполагается, что отдельные признаки (случаи или элементы временного ряда) являются независимыми; 2) генерируемые кластеры имеют различные профили (за счет искусственного разбиения признаков на несколько интервалов, значения внутри каждого из которых сгенерированы согласно индивидуальным вероятностным распределениям); 3) размеры кластеров варьируются от 10 до 100 элементов; 4) генерируются несколько кластеров шумов для проверки способности алгоритма выделять и исключать из рассмотрения незначимые генные кластеры.

Использованный для тестирования сгенерированный набор данных состоял из четырех профилей генных кластеров (табл. 1). Каждый кластер содержал различное количество генов, всего 85 информативных генов. Случайным образом согласно нормальному распределению генерировались дополнительно 50 генов, которые были распределены по 100 объектам согласно $N(0,1)$ и представляли собой шум в данных. Все признаки были разделены на пять классов (по 20 объектов каждого класса).

Таблица 1

Параметры сгенерированного набора данных

Кластер	Средние значения профилей генов для пяти классов	Стандартное отклонение	Количество генов
1	1, –1, –1, –1, –1	0,2	5
2	–1, 1, –1, –1, –1	0,2	10
3	–1, –1, 1, –1, –1	0,2	20
4	–1, –1, –1, 1, –1	0,2	50

В результате применения метода были выявлены все четыре функциональных модуля генов и отсеян шум. На рис. 3 представлены четыре профиля полученных кластеров.

Таким образом, экспериментально удалось показать, что предложенный метод способен выявлять кластеры, имеющие разные размеры, и отфильтровывать неинформативные признаки. Rand-индекс [11] качества кластеризации равен единице.

Для выполнения сравнительного анализа результатов, полученных предложенным методом и методом *k*-средних, был использован R-пакет cclust, который позволяет итерационно смещать центры кластеров к средним значениям областей соответствующей диаграммы Вороного. В качестве расстояния между объектами используется евклидово расстояние. Основным недостатком метода *k*-средних является необходимость изначального определения числа кластеров, которое в большинстве случаев заранее неизвестно. Для проведения эксперимента количество кластеров было определено равным пяти в предположении, что все неинформативные гены определяют отдельный кластер.

В результате метод *k*-средних оказался неспособным выявить все информативные кластеры. Было получено пять кластеров, из который кластер 1 являлся объединением исходных кластеров 1 и 2 (см. табл. 1), кластер 2 соответствовал исходному кластеру 3, кластер 3 соответствовал исходному кластеру 4 и неинформативные гены были разбиты на два оставшихся кластера (табл. 2). Rand-индекс качества кластеризации равен 0,79.

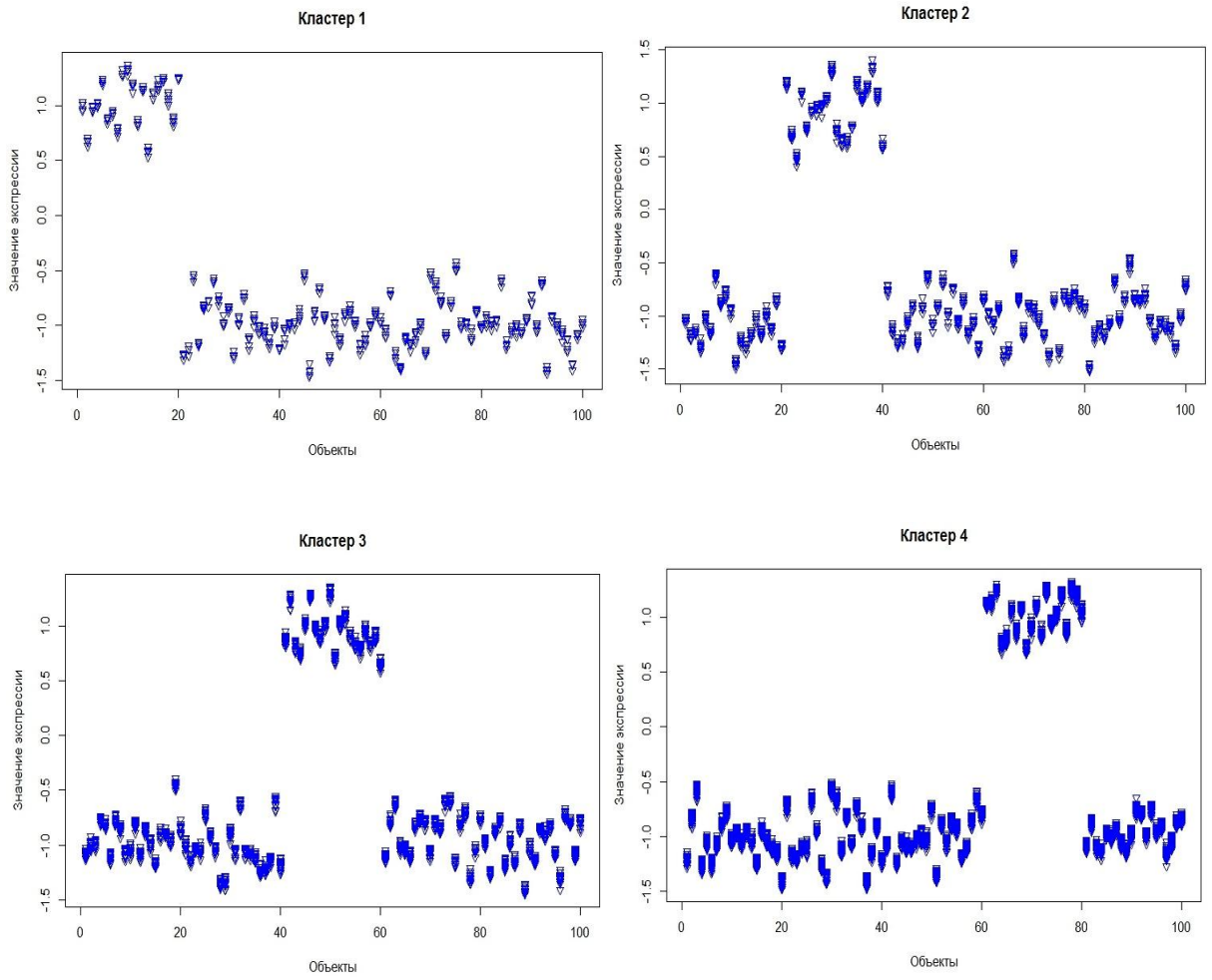


Рис. 3. Профили кластеров, полученных с использованием предложенного метода

Таблица 2

Результаты кластеризации методом *k*-средних

Кластер	Значения генов исходного набора данных	Кластеры генов исходного набора данных	Количество генов
1	1–15	1, 2	15
2	16–35	3	20
3	36–85	4	50
4	86–96, 100, 103–105, 107, 109, 111–113, 117, 118, 122, 123, 125, 126, 128, 130	Неинформативные гены	28
5	97–99, 101, 102, 106, 108, 110, 114–116, 119–121, 124, 127, 129, 131–135	Неинформативные гены	22

Второй сравнительный эксперимент на том же наборе данных был проведен с использованием гибридного кластерного метода hybridHclust [12], представляющего собой иерархический метод кластеризации, широко применяемый для анализа данных генной экспрессии и позволяющий выявить иерархическую организацию кластерной структуры. Метод позволяет компенсировать ряд недостатков базовых методов иерархической кластеризации, объединяя кластеризацию снизу вверх с кластеризацией по нисходящему принципу. Для проведения эксперимента использовался R-пакет hybridHclust, а для вычисления близости объектов приме-

нялось евклидово расстояние. В результате было определено 14 кластеров, среди которых имеются все четыре исходных кластера из табл. 1. В кластер 4 кроме исходного кластера вошел и неинформативный ген 113. Кроме того, были получены 10 дополнительных кластеров, которые являются неинформативными в связи с тем, что соответствующие им гены имеют случайным образом сгенерированный профиль (табл. 3).

Таблица 3

Результаты кластеризации методом hybridHclust

Кластер	Значения генов исходного набора данных	Кластер	Значения генов исходного набора данных
1	1–5	8	90, 96
2	6–15	9	92, 134
3	16–353	10	104, 111
4	36–85, 113	11	107, 130
5	94, 114	12	89, 109
6	100, 101	13	102, 110
7	105, 133	14	112, 131

Можно сделать вывод, что метод hybridHclust в отличие от предложенного не позволяет точно определить исходные кластеры и отделить шум от информативных признаков.

5. Результаты тестирования метода на наборе данных по анализу клеточного цикла дрожжевых грибков

Набор данных дрожжевых грибков [9] является результатом применения технологии микрочипов для анализа клеточного цикла. Набор состоит из значений экспрессии 2500 генов, полученных в результате 79 экспериментов. Дрожжи представляют собой классический модельный организм как для исследований фундаментальных генетических проблем, так и для изучения наследственных болезней человека в связи с тем, что многие из основных клеточных процессов консервативны и одинаковы у дрожжей и у человека. Например, многие из известных сейчас молекул и метаболических путей, вовлеченных в формирование злокачественных новообразований (например, клеточный цикл и его контролирование, репарация ДНК и т. д.), впервые были открыты и исследованы на дрожжах. Набор данных содержит результаты экспериментов с различными штаммами грибов для выявления генов, ассоциированных с различными этапами клеточного цикла.

С помощью предложенного метода из набора данных были выделены 20 функциональных модулей, каждый из которых характеризуется множеством генов со схожими профилями экспрессии. В состав каждого модуля вошли гены, имеющие наибольшие весовые коэффициенты, соответствующие компонентам значимых собственных векторов. Изначально предполагалось, что гены, относящиеся к одному модулю, вовлечены в похожие биологические процессы. Для определения биологических процессов для генных кластеров была использована веб-среда YeastMine [13], которая позволяет получить генетическую информацию для дрожжевых грибов Saccharomyces cerevisiae. Ее применение обеспечивает возможность поиска по набору генов степени их вовлечения в биологический процесс, сигнальный путь, GO-аннотацию генов и т. д. В табл. 4 представлен ряд выделенных функциональных модулей, которые соответствуют процессам биогенезиса протеинов, репликации и восстановления ДНК, энергетического метаболизма, деградации протеинов, сворачивания протеинов, а также протеинам теплового шока, циклу Кребса, механизму разложения аллантоина и регуляции гистонов.

Согласно разработанной процедуре итерационной кластеризации модули, состоящие из большого числа генов, были повторно кластеризованы для выявления подгрупп в составе модуля. Для наибольшей группы, содержащей 230 генов, были выделены два подмодуля, соответствующие процессам гликолиза и клеточного цикла. Таким образом удалось показать, что результаты применения метода позволяют выявить функциональные модули генов, являющиеся информативными и с биологической точки зрения.

Таблица 4

Пример выявленных функциональных модулей генов, аннотированных согласно YeastMine

Функциональный модуль	Гены	Функциональный модуль	Гены
Биогенезис протеинов	RPL8B, RPL34B, RPS19A, RPS26B, RPL7B, RPL14A, RPS6B, RPL33B, RPS1B, RPL9A, RPL11B, RPS24A, RPL26B, RPS15, RPL24A, RPL12B, RPS0A, RPS19B, RPL19B, RPL27B, RPS29B, RPL13A, RPS4B, RPL33A, RPS29A, RPL18B, RPL11A,...	Протеины теплового шока	SSA2, SSA1, ECM10, SSA4, KAR2, STI1, SIS1, SUR2, YME1, SRP21, FCY1, SUI1, VAM3, PFD4
		Регуляция гистонов	HHT1, HTA1, HTB1, HHF2, HTA2, HHT2, HHF1, HTB2, HHO1
Репликация и восстановление ДНК	CDC45, KIM2, MSH2, DUN1, POL30, SMC3, RFA2, RNR1, SWE1, POL2, POL32, RFA1, SEN34, PIF1, SPH1, OGG1, CDC21, RHC18, RFC3, RAD27, TUB4, ZDS2, ASF1, HCM1, MSH6, RFC5, SAS2, BNI4, CDC2, PMS1, ARP1, POL1, TOP1, RAD51, FCP1, PMT5, PCH1, RNR3, MRE11, ALG2, PRI1, RFA3, BUD2, DUT1, ASF2, ADK2, CDC9, ECM25, RFC4	Цикл Кребса	CYT1, SDH2, MDH1, SDH4, QCR6, ACH1, COR1, SDH3, RIP1
		Механизм разложения аллантоина	DAL2, DAL1, DAL3, DAL5, MEP2
Деградация протеинов	SCL1, PUP2, PRE2, PRE5, PRE1, PRE3, RPN6, PRE4, RPT6, PRE9, RPT1, RPN9, RPT4, RPN10, RPN11, PRE7, PUP1, RPN3, PRE8, PRE10, UFD1, UMP1, PRE6, RPN12, PUP3, RPN7, RPT3, RPN2, RPT5, GSH2, QRI8, STE24	Энергетический метаболизм	MRPL9, MRPL13, ATP12, PET123, COX12, MRPL35, MRP2, ATP11, MRP17, DBI56, QCR8, MRP51, PPA2, HXT4, COX17, RML2, MRPS5, MRPL38, MRP7, MRPL25, ATP14, MSS51, FMC1, MSM1, MRPL8, CYT2, COX14, MEF1, MRPL32, MSN4, CBP4, PHB1, MSF1, MRPL16, MRPL6, IMG2, MRP13, NAM2, MAS1, MRPL36, NAM9, CYC3, KIM4, CBP3, STS1, COX10, MBA1, MRPL31, ECM19, STP4, CAF16, COX13, CAP1

Для сравнения набор данных по анализу клеточного цикла дрожжевых грибков был кластеризован с использованием метода *k*-средних. Начальное количество кластеров было задано равным 20, что соответствовало результату, полученному с использованием предложенного метода. Метод *k*-средних позволил определить группу генов, связанных с процессом биогенеза протеинов, однако даже эта группа была определена не полностью и часть связанных с данным процессом генов была перераспределена в другие кластеры. Кроме того, метод *k*-средних был не способен выявить малые функциональные модули, такие как гистоны и цикл Кребса, так как целевая функция данного метода предполагает построение кластеров сравнимых размеров и малые кластеры при этом всегда являются составной частью кластеров большего размера.

Набор данных по анализу клеточного цикла дрожжевых грибков также был кластеризован с помощью метода hybridHclust [12]. Полученная в результате вычислительного эксперимента дендрограмма оказалась трудна в интерпретации из-за большого числа анализируемых генов. Размерность большинства полученных кластеров была очень мала – в диапазоне от двух до восьми элементов. Надо отметить, что ряд кластеров, выделенных с помощью метода hybridHclust, были биологически интерпретируемы. Например, пять генов, отвечающих за гликолиз, выделены в один кластер; восемь из девяти генов, связанных с регуляцией гистонов, также образовали кластер. Однако при этом оказалось невозможно выделить кластеры больших размеров. Кроме того, полученные кластеры были чувствительны к малым вариациям в данных, что приводило к изменению их состава.

Заключение

Согласно разработанному методу построения кластеров генетических данных с использованием теории рандомизированных матриц первоначально рассчитываются главные компоненты исходной корреляционной матрицы генных профилей. Для учета дисбаланса между количеством генов и экспериментов, полученных по технологии микрочипов, в качестве конечных выбираются главные компоненты, которые соответствуют ограниченному количеству собственных значений. Величины этих собственных значений значимо отличаются от значений, полученных при анализе рандомизированной корреляционной матрицы R. Статистические характеристики случайной корреляционной матрицы позволяют отфильтровать шум из исходной анализируемой матрицы данных и выделить статистически значимые функциональные модули.

Сравнительный анализ результатов применения предложенного метода и двух известных методов (k-средних и иерархической кластеризации hybridHclust) показал преимущество предложенного метода выделения информативных кластеров. Кластеризация методом k-средних не способна была распознать кластеры малых размеров, и, наоборот, результатом иерархической кластеризации стало большое количество кластеров малого размера, некоторые из которых являлись составной частью реальных больших кластеров. Кроме этого, метод hybridHclust оказался неспособным отфильтровать неинформативные гены.

Таким образом, сравнительный анализ разработанного метода построения кластеров генетических данных с использованием рандомизированных матриц с наиболее известными методами-аналогами показал, что он позволяет более корректно выявлять структуру в данных генной экспрессии как для больших, так и для малых кластеров, успешно фильтруя при этом неинформативные гены.

Список литературы

1. Liang, S. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures / S. Liang, S. Fuhrman, R. Somogyi // Pacific Symp. on Biocomputing (PSB’98). – Hawaii, 1998. – Vol. 3. – P. 18–29.

2. Cluster analysis and display of genome-wide expression patterns / M.B. Eisen [et al.] // Proceedings of the National Academy of Sciences of the United States of America. – 1998. – Vol. 95. – P. 14863–14868.

3. Analysis of gene expression data using self-organizing maps / P. Toronen [et al.] // FEBS Letters. – 1999. – Vol. 451. – P. 142–146.

4. The R Project for Statistical Computing. R Foundation for Statistical Computing [Electronic resource]. – 2009. – Mode of access : http://www.R-project.org. – Date of access : 10.09.2015.

5. Bioconductor case studies / F. Hahne [et al.]. – Springer Science & Business Media, 2010. – 296 p.

6. Cluster – Cluster analysis and visualization software [Electronic resource]. – 2015. – Mode of access : http://rana.lbl.gov/EisenSoftware.htm. – Date of access : 19.08.2015.

7. Cyber-T – microarray analysis web interface from UCI's Institute for Genomics and Bioinformatics [Electronic resource]. – 2015. – Mode of access : http://cybert.microarray.ics.uci.edu. – Date of access : 16.09.2015.

8. SNOMAD – Standardization and normalization of microarray data [Electronic resource]. – 2015. – Mode of access : http://pevsnerlab.kennedykrieger.org/snomadinput.html. – Date of access : 12.09.2015.

9. Yeast cell cycle analysis project [Electronic resource]. – 2015. – Mode of access : http://genome-www.stanford.edu/cellcycle. – Date of access : 10.04.2015.

10. Varimax – rotation methods for factor analysis [Electronic resource]. – 2015. – Mode of access : https://stat.ethz.ch/R-manual/R-devel/library/stats/html/varimax.html. – Date of access : 17.09.2015.

11. Morey, L.C. The measurement of classification agreement: an adjustment to the rand statistic for chance agreement / L.C. Morey, A. Agresti // Educational and Psychological Measurement. – 1984. – Vol. 44. – P. 33–37.

12. Chipman, H. Hybrid hierarchical clustering with applications to microarray data / H. Chipman, R. Tibshirani // Biostatistics. – 2006. – Vol. 7, № 2. – P. 286–301.

13. YeastMine: saccharomyces genome database [Electronic resource]. – 2015. – Mode of access : http://yeastmine.yeastgenome.org/yeastmine/begin.do. – Date of access : 06.09.2015.

Поступила 09.10.2015

Объединенный институт проблем
информатики НАН Беларуси,
Минск, Сурганова, 6
e-mail: novosel@newman.bas-net.by

N.A. Novoselova, I.E. Tom

METHOD OF CONSTRUCTION OF GENETIC DATA CLUSTERS

The paper presents a method of construction of genetic data clusters (functional modules) using the randomized matrices. To build the functional modules the selection and analysis of the eigenvalues of the gene profiles correlation matrix is performed. The principal components, corresponding to the eigenvalues, which are significantly different from those obtained for the randomly generated correlation matrix, are used for the analysis. Each selected principal component forms gene cluster. In a comparative experiment with the analogs the proposed method shows the advantage in allocating statistically significant different-sized clusters, the ability to filter non-informative genes and to extract the biologically interpretable functional modules matching the real data structure.