Данные о треках и плейлистах будут храниться в реляционной базе данных. Реляционные базы данных основаны на реляционной модели – интуитивно понятном, наглядном табличном способе представления данных. Каждая строка, содержащая в таблице такой базы данных, представляет собой запись с уникальным идентификатором, который называют ключом. Столбцы таблицы имеют атрибуты данных, а каждая запись обычно содержит значение для каждого атрибута, что дает возможность легко устанавливать взаимосвязь между элементами данных. Для этого будет использована MySQL. MySQL – свободная реляционная система управления базами данных.

Итак, разрабатываемое приложение может стать удобным и востребованным сервисом для людей, нуждающихся в психологической помощи. Для его создания используются современные информационные технологи+и. В разработке сервиса участвует группа специалистов, включающая, кроме программистов, психолога и музыкального работника. Для тестирования системы будут привлечены студенты и преподаватели медицинского университета.

АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ РЕЗУЛЬТАТОВ АНАЛИЗА КРИВЫХ ПЛАВЛЕНИЯ С ВЫСОКИМ РАЗРЕШЕНИЕМ

ALGORITHMS FOR RESULTS' CLUSTERING OF MELTING CURVES' ANALYSIS WITH HIGH RESOLUTION

Ю. И. Белькович[1,2], Е. В. Снытков[1,2], Б. А. Тонконогов[1,2]

Y. I. Bel'kovich[1,2], E. V. Snitkov[1,2], B. A. Tonkonogov[1,2]

[1]Белорусский государственный университет, БГУ, г. Минск, Республика Беларусь
[2]Учреждение образования «Международный государственный экологический институт имени А. Д. Сахарова» Белорусского государственного университета, МГЭИ им. А. Д. Сахарова БГУ, г. Минск, Республика Беларусь
boristonkonogov@iseu.by, lily.belkovich@gmail.com

[1]Belarusian State University, BSU, Minsk, Republic of Belarus
[2]International Sakharov Environmental Institute of Belarusian State University, ISEI BSU Minsk, Republic of Belarus

Рассмотрены основные принципы кластеризации данных, описаны этапы ее проведения и способы определения схожести объектов и интерпретации полученных результатов и дана характеристика различным способам проведения плавления ДНК.

Basic principles of data clustering are considered, stages of its implementation and methods to determine object similarity and result interpretation are described and characteristics of various conduct methods for DNA melting are given.

Ключевые слова: алгоритмы кластеризации, кривые плавления, высокое разрешение.

Keywords: clustering algorithms, melting curves, high resolution.

https://doi.org/10.46646/SAKH-2022-2-416-419

Введение. Плавление с высоким разрешением (High Resolution Melting (HRM) – это новый гомогенный метод пост-ПЦР (полимеразной цепной реакции) в закрытой пробирке, позволяющий исследователям геномов анализировать генетические вариации (однонуклеотидный полиморфизм (Single Nucleotide Polymorphism (SNP), мутации, метилирование) в ампликонах (фрагментах) ДНК (дезоксирибонуклеиновой кислоты) и РНК (рибонуклеиновой кислоты). Он выходит за рамки классического анализа кривой плавления, позволяя изучать термическую денатурацию двухцепочечной ДНК гораздо более подробно и с гораздо более высоким информационным выходом, чем когда-либо прежде. Данные, полученные вследствие HRM-анализа, далее интерпретируются для получения генотипов исследованных образцов. Чтобы получить более достоверные результаты существует возможность прибегнуть к кластеризации полученных данных наряду с интерпретацией их с помощью специализированных пакетов программного обеспечения, разрабатываемых, как правило, производителями оборудования для HRM-анализа.

Производители амплификаторов для детектирования результатов ПЦР, как правило, разрабатывают оригинальное программное обеспечение для HRM-анализа, например:

– Precision Melt Analysis Software (Bio-Rad);
– High Resolution Melt Software (Applied Biosystems ThermoFisher Scientific);
– The LightCycler* 480 Gene Scanning Software (Roshe).

Имеются альтернативные варианты программного обеспечения с открытым исходным кодом, реализующего определенный алгоритм кластеризации и выполняющего схожие функции, например:

– Python-HRM;

– uAnalyze;

– Novallele HRM Analyzer.

В данный момент на базе учреждения образования «Международный государственный экологический институт имени А. Д. Сахарова» Белорусского государственного университета (МГЭИ им. А. Д. Сахарова БГУ) ведется разработка программного обеспечения, реализующего несколько алгоритмов кластеризации, и позволяющего провести соответствующий анализ с помощью любого из них.

Алгоритмы кластеризации. Кластеризация (кластерный анализ) – задача группировки множества объектов на подмножества (кластеры) таким образом, чтобы объекты из одного кластера были более похожи друг на друга, чем на объекты из других кластеров по какому-либо критерию. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма.

Типичный процесс кластеризации включает следующие этапы:

1) представление данных (количество классов, тип и масштаб функций, доступных для алгоритма кластеризации);

2) определение меры близости паттерна (шаблона) соответствующей предметной области;

3) кластеризация или группировка (может быть четкой (данные разделяются на группы) иди нечеткой (каждый элемент имеет переменную степень принадлежности к каждому из выходных кластеров);

4) абстракция данных (при необходимости компактное описание каждого кластера обычно в терминах прототипов кластеров или репрезентативных шаблонов);

5) оценка результатов (при необходимости).

Все алгоритмы кластеризации при представлении данных будут создавать кластеры независимо от того, содержат ли эти кластеры данные или нет. Если кластеры содержат данные, то некоторые алгоритмы кластеризации могут получить «лучшие» кластеры, чем другие. Таким образом, оценка результатов процедуры кластеризации имеет несколько аспектов. Один из них – это оценка предметной области, но не самого алгоритма кластеризации. Данные, которые не содержат кластеров, не должны обрабатываться алгоритмом кластеризации. Анализ валидности кластеров, напротив, представляет собой оценку результатов процедуры кластеризации. Часто в этом анализе используется определенный критерий оптимальности, обычно определяющийся субъективно. Структура кластеризации действительна, если она не может возникнуть случайно или как артефакт алгоритма кластеризации. Когда используются статистические подходы к кластеризации, то анализ выполняется путем применения статистических методов и проверки гипотез.

Существует 3 типа валидационных исследований:

1) внешняя оценка достоверности – сравнивает восстановленную структуру с априорной структурой;

2) внутренняя проверка достоверности – определяет, действительно ли структура соответствует данным;

3) относительный тест – сравнивает две структуры и измеряет их относительные достоинства.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата.

Для определения «похожести» объектов нужно составить вектор характеристик для каждого объекта – как правило, это набор числовых значений, к примеру, рост или вес человека. Однако существуют также алгоритмы, которые работают и с качественными, так называемыми категорийными, характеристиками. После того, как определен вектор характеристик, проводится нормализация, чтобы все компоненты давали одинаковый вклад при расчете так называемого «расстояния». В процессе нормализации все значения приводятся к некоторому диапазону, например, [-1, -1] или [0, 1]. Наконец, для каждой пары объектов измеряется «расстояние» между ними – степень похожести. Существует множество различных метрик, выбор которых полностью лежит на исследователе, поскольку результаты кластеризации могут существенно отличаться при использовании разных мер. Некоторыми из указанных метрик являются:

1. Евклидово расстояние. Самая распространенная функция расстояния, которая представляется геометрическим расстоянием в многомерном пространстве:

$$\rho(x, x') = \sqrt{\sum_{i}^{n} (x_i - x_i)^2}. \quad (1)$$

2. Расстояние городских кварталов (манхэттенское расстояние). Это расстояние является средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако для этой меры влияние отдельных больших разностей (выбросов) уменьшается, так как они не возводятся в квадрат:

$$\rho(x, x') = \sum_{i}^{n} |x_i - x_i|^2. \quad (2)$$

3. Расстояние Чебышева. Это расстояние может оказаться полезным, когда нужно определить два объекта как «различные», если они различаются по какой-либо одной координате:

$$\rho(x, x') = ma(|x_i - x_i|). \quad (3)$$

4. Степенное расстояние. Применяется в случае, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются:

$$\rho(x, x') = \sqrt[r]{\sum_{i}^{n} (x_i - x_i)^p}, \quad (4)$$

где r и p – параметры, определяемые пользователем. Параметр p ответственен за постепенное взвешивание разностей по отдельным координатам, а параметр r – за прогрессивное взвешивание больших расстояний между объектами. Если оба указанные параметра равны 2, то это расстояние совпадает с расстоянием Евклида.

Методы кластеризации классифицируются по:

1) способу обработки данных:

– иерархические:

• агломеративные (CURE, ROCK, CHAMELEON);

• дивизимные (BIRCH, MST);

– неиерархические методы:

• итеративные (k-средних, PAM, CLOPE, LargeItem);

2) способу анализа данных:

– четкие;

– нечеткие.

3) количеству применений:

– одноэтапные;

– многоэтапные (рис. 1).

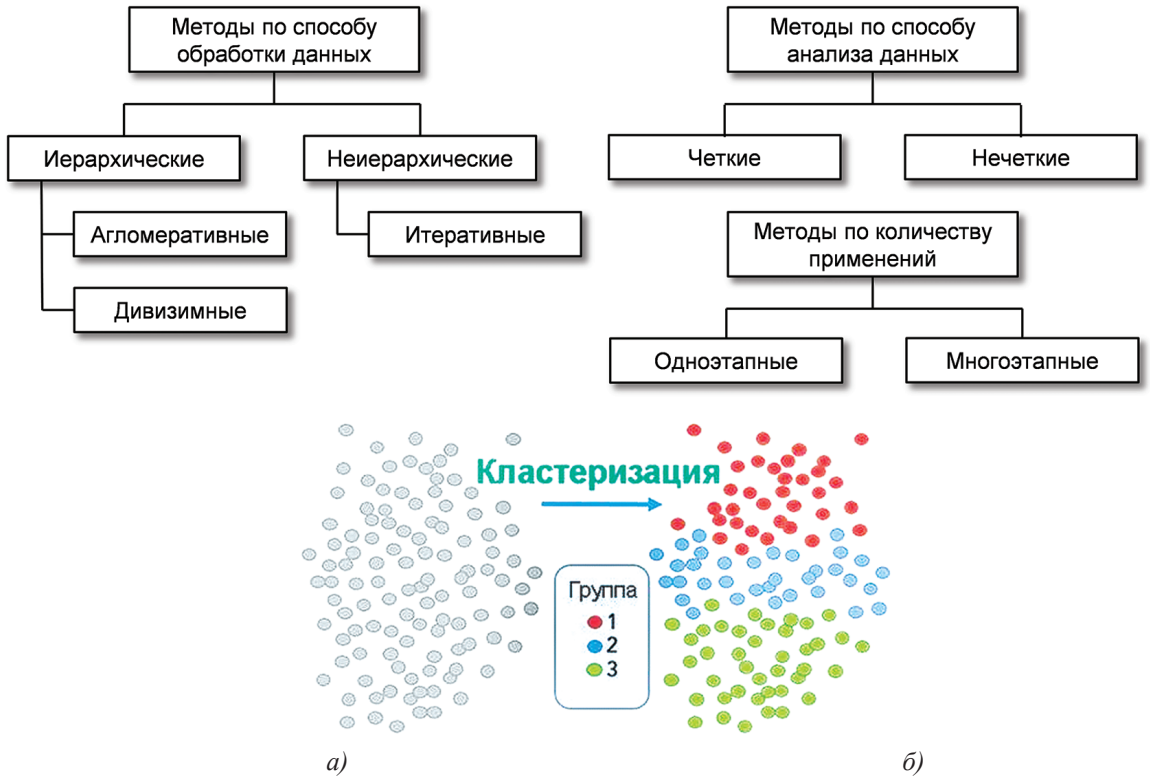

Рисунок 1 – Обобщенная классификация (а) и визуальная интерпретация (б) кластерных методов, характерные для большинства научно-исследовательских задач

В зависимости от используемых методов на выходе формируются соответствующие данные. Например, иерархические алгоритмы на выходе предоставляют некую иерархию кластеров, и можно выбрать любой уровень этой иерархии для того, чтобы интерпретировать результаты алгоритма. Неиерархические – это, фактически, все алгоритмы, которые на выходе иерархию не предоставляют (или выбор интерпретации происходит не по уровню иерархии).

Процесс плавления ДНК. Анализ кривой плавления представляет собой оценку характеристик диссоциации двухцепочечной ДНК при нагревании. При повышении температуры двойная нить начинает диссоциировать, что приводит к увеличению интенсивности поглощения и гиперхромности. Температура, при которой денатурируется 50 % ДНК, называется температурой плавления.

Собранная информация может быть использована для определения наличия и идентичности однонуклеотидных полиморфизмов (SNP). Это связано с тем, что пары оснований GC имеют 3 водородные связи между собой,

в то время как пары оснований АТ имеют только 2. ДНК с более высоким содержанием GC из-за ее источника или из-за SNP будет иметь более высокую температуру плавления, чем ДНК с более высоким содержанием АТ.

Эта информация также дает важные сведения о способе взаимодействия молекулы с ДНК. Молекулы, такие как интеркаляторы, располагаются между парами оснований и взаимодействуют посредством укладки. Это оказывает стабилизирующее действие на структуру ДНК, что приводит к повышению температуры ее плавления. Точно так же увеличение концентрации солей помогает рассеять отрицательное отталкивание между фосфатами в остове ДНК. Это также приводит к повышению температуры ее плавления. И наоборот, pH может оказывать негативное влияние на стабильность ДНК, что может привести к снижению ее температуры плавления.

Анализ кривой плавления имеет несколько ограничений. Абсолютное положение и ширина кривых плавления зависят от концентрации красителя, температуры и скорости перехода. Добавление интеркаляторов, таких, как бромид этидия, повышает температуру плавления и уширяет переход плавления.

Заключение. Таким образом, высокочувствительный анализ кривых плавления – это относительно новый метод анализа, позволяющий выявлять новые варианты генов, делать скрининг образцов ДНК на однонуклеотидные полиморфизмы, выявлять вставки и делеции и другие неизвестные мутации и определять процент метилированной ДНК в образцах. Однако при анализе гетерозиготных вариантов кривые плавления могут быть трудно различимы, поэтому в таких случаях для анализа используют другие графические зависимости, такие, как графики пиков плавления и сравнения кривых плавления.

Точность алгоритмов зависит от многих факторов, поэтому при анализе результатов кривых плавления не стоит полагаться лишь на один вариант детектирования. В рамках научных исследований в МГЭИ им. А. Д. Сахарова БГУ ведется разработка программных модулей, реализующих некоторые алгоритмы кластеризации, и позволяющих с их помощью провести соответствующий анализ [1–3].

ЛИТЕРАТУРА

1. *Jain, A.* Data Clustering: A Review / A. Jain, M. Murty, P. Flynn // ACM Computing Surveys, 1999.31(3). P. 264–323.

2. *Ririe, K.* Product Differentiation by Analysis of DNA Melting Curves during the Polymerase Chain Reaction / K. Ririe, R. Rasmussen, C. Witter // Anal. Biochem. 1997.252(2). P. 857–859.

3. *Wittwer, C. T.* High-resolution DNA melting analysis: advancements and limitations / C. T. Witter // Human Mutation, 2009.30(6). P. 857–859.

ПРОЕКТ ПРОГРАММНОЙ СИСТЕМЫ ДЛЯ АНАЛИЗА ВЗАИМОДЕЙСТВИЯ БЕЛКОВ В УСЛОВИЯХ СЛАБОГО СТРУКТУРНОГО ПОДОБИЯ

SOFTWARE SYSTEM PROJECT FOR PROTEINS' INTERACTION ANALYSIS UNDER CONDITIONS OF WEAK STRUCTURAL SIMILARITY

А. Д. Казмерчук[1,2], Е. В. Снытков[1,2], Б. А. Тонконогов[1,2]

A. D. Kazmerchuk[1,2], E. V. Snytkov[1,2], B. A. Tonkonogov[1,2]

[1]Белорусский государственный университет, БГУ
г. Минск, Республика Беларусь
[2]Учреждение образования «Международный государственный экологический институт имени А. Д. Сахарова» Белорусского государственного университета, МГЭИ им. А. Д. Сахарова БГУ
г. Минск, Республика Беларусь
antkaz566@gmail.com, evgeni.snytkov@iseu.by

[1]Belarusian State University, BSU
Minsk, Republic of Belarus
[2]International Sakharov Environmental Institute of Belarusian State University, ISEI BSU
Minsk, Republic of Belarus

Рассмотрены практическое назначение, алгоритм функционирования, архитектура, технологии и средства реализации, функциональность и тестирование проекта программной системы для анализа взаимодействия белков в условиях слабого структурного подобия, базирующейся на биоинформационных технологиях, реализующей расчетные методы и математические модели на различных вычислительных уровнях и позволяющей принимать исследовательские решения в различных областях биологии и медицины.

Practical purpose, functioning algorithm, architecture, technologies and means of implementation, functionality and testing of software system project for proteins' interaction analysis under conditions of weak structural similarity, based on bioinformatics technologies, implementing computational methods and mathematical models at