Использование вероятностного подхода для анализа последовательностей аминокислот

Штефан Михаил, Боровитинов Илья

Лицей №1533(информационных технологий)
11 класс
Научный руководитель: Миронов Андрей Александрович, д.б.н., профессор, преподаватель факультета биоинформатики и биоинженерии МГУ им. М.В. Ломоносова

Одной из основных отраслей современной биоинформатики является анализ и исследование последовательностей ДНК. Большую роль в этой области играют алгоритмы, позволяющие сравнивать и анализировать схожести между двумя аминокислотными последовательностями различных организмов.

На основе полученной информации, специалисты получают возможность предсказывать функции, структуру белков, а так же исследовать их эволюцию. Это позволяет понимать происхождение схожестей двух организмов, корректировать предложенные ранее цепочки эволюции, находить качества и уязвимости еще не исследованных организмов.

Классическим методом, используемым для анализа и сравнения последовательностей, является выравнивание последовательностей. Результатом его является расположение двух аминокислотных последовательностей, представленных в виде символьного ряда, друг над другом таким образом, чтобы наглядно были видны места схожести и различия (схожие участки последовательностей стоят друг над другом, несхожие – нет):

```
Sequence 1: ABVSFRQEDA     Alignment:         ABVSFRQ-EDA---
Sequence 2: FRQSADASVX                        ----FRQSADASVX
```

Для двух последовательностей можно провести большое количество различных выравниваний, при этом разница между ними будет в количестве демонстрируемых схожестей/различий (*весе выравнивания*). Поэтому часто используются алгоритмы, ищущие выравнивание, при котором вес максимален для двух данных последовательностей (например, опубликованный в 1970 г. алгоритм Нидлмана-Вунша [http://www.sciencedirect.com/science/article/pii/0022283670900574])

На практике выясняется, что принцип «лучше больше, да лучше» (когда алгоритмы стремятся найти выравнивание, при котором количество показанных схожестей максимально) не всегда является оправданным, особенно в случаях, когда требуется сравнить сильно различающиеся последовательности. Причина неточностей работы алгоритмов, а также главная сложность в построении алгоритмов анализа последовательностей заключается в том, что исследователю необходимо сложнейшую и не до конца изученную систему, построенную природой, выразить в виде нескольких простых (сравнительно) формул. Одним из оптимальных решений этой проблемы, которое реализовано и проверено в данном проекте, является анализ всех возможных выравниваний посредством вероятностных методов и представление результатов в наглядной форме.

В реализованном проекте предложены первая реализация алгоритма нахождения вероятностей совпадения пары аминокислот при локальном выравнивании последовательностей, а также метод визуализации получаемых результатов.

Краткое описание алгоритма:

- Для каждой пары аминокислот в последовательностях посредством анализа всех возможных выравниваний, проводимых с данными последовательностями, определяется вероятность их совпадения при произвольном выравнивании.
- По завершении, полученная матрица выводится в виде изображения, в котором цвет каждого пикселя отвечает за вероятность выравнивания пары нуклеотидов.
- Алгоритм создает ориентированный граф, любой путь из начала в конец которого символизирует выравнивание двух заданных последовательностей, и количество всех таких путей равно количеству возможных выравниваний.
- После этого для каждой вершины графа вычисляется *качество* - статистическая сумма всех возможных весов выравниваний, которые к ней можно провести как для прямого, так и обратного хода по графу.
- Затем граф обрабатывается алгоритмами, вычисляющими для каждой вершины статистическую сумму всех возможных весов выравниваний, которые к ней можно

провести как для прямого (*при помощи алгоритма просмотра вперед*), так и обратного (*алгоритм просмотра назад*) хода по графу

Для пары аминокислот i,j вероятность совпадения пары аминокислот вычисляется по приведенной ниже формуле, где Z+ - результат алгоритма просмотра вперед, Z- - результат просмотра назад, l1 и l2 — длины последовательностей.

$$P(i,j) = \frac{Z+(i,j) * Z-(i,j)}{Z(l1,l2)}$$

Результатом данной работы является программа с Web-интерфейсом, реализующая:

- Алгоритм нахождения вероятностей при локальном выравнивании последовательностей.
- Алгоритм выравнивания последовательности относительно двух нуклеотидов либо локально, либо глобально.

Правильность работы данного метода была проверена на задаче нахождения расстояний и восстановления филогенетического древа по результатам анализа.

Проект реализован на платформе Eclipse на языке Java с использованием библиотек Google Web Tools для реализации Web-интерфейса.

Литература

1. Durbin R., Eddy S.R., Krogh A., Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids // Cambridge University Press, 1998.
2. Mount D. Bioinformatics: Sequence and Genome Analysis // Cold Spring Harbor Laboratory Press, 2004.
3. Lesk A. Introduction to Bioinformatics // Oxford University Press, USA, 2008.
4. Cormen T.H., Leiserson C.E., Rivest R.L., Stein C. Introduction to Algorithms, 2nd Edition // MIT Press, USA, 2009
5. Dai N., Mandel L., Ryman A. Eclipse Web Tools Platform: Developing Java Web Applications // Pearson Education, 2007