

Библиотека Чебышевского сборника

Министерство науки и высшего образования Российской Федерации
Министерство просвещения Российской Федерации
Российская академия наук
Московский государственный университет им. М. В. Ломоносова
Математический институт им. В. А. Стеклова РАН
Санкт-Петербургское отделение Математического института им. В. А. Стеклова РАН
Санкт-Петербургский государственный университет
Институт истории естествознания и техники им. С. И. Вавилова РАН
Московский педагогический государственный университет
Тульский государственный педагогический университет им. Л. Н. Толстого
Математический центр мирового уровня МИАН
Тульский государственный университет

**Алгебра, теория чисел, дискретная геометрия
и многомасштабное моделирование: современные
проблемы, приложения и проблемы истории**

**Материалы XXIV Международной конференции,
посвящённой 110-летию со дня рождения академика
Юрия Владимировича Линника
и 110-летию со дня рождения профессора
Андрея Борисовича Шидловского
и 80-летию со дня рождения профессора
Геннадия Ивановича Архипова**

Тула, 14–17 мая 2025 года

ББК 22.1
УДК 51
А45

Председатель программного комитета — профессор В. Н. Чубариков

Сопредседатели программного комитета:

член-корреспондент В. М. Бухштабер;

академик С. В. Конягин;

академик Ю. В. Матиясевич;

академик В. П. Платонов

Ответственный секретарь — Н. М. Добровольский

Программный комитет: Балаба И. Н. (Тула), Берник В. И. (Минск, Белоруссия), Быковский В. А. (Хабаровск), Востоков С. В. (Санкт-Петербург), Всемирнов М. А. (Санкт-Петербург), Гашков С. Б. (Москва), Гриценко С. А. (Москва), Деза Е. И. (Москва), Демидов С. С. (Москва), Долбилин Н. П. (Москва), Зубков А. М. (Москва), Иванов А. О. (Москва), Иванов В. И. (Тула), Королёв М. А. (Москва), Кузнецов В. Н. (Саратов), Лауринчикас А. (Вильнюс, Литва), Михалёв А. В. (Москва), Мищенко С. П. (Ульяновск), Мороз Б. З. (Москва), Нестеренко Ю. В. (Москва), Нижников А. И. (Москва), Ольшанский А. Ю. (Нашвилл, США), Пачев У. М. (Нальчик), Рахмонов З. Х. (Душанбе, Таджикистан), Семёнов А. Л. (Москва), Устинов А. В. (Хабаровск), Фомин А. А. (Москва), Чирский В. Г. (Москва), Antonio Mucherino (France), Patrick Sole (France), Mathieu Dutour (France), Aleksandar Jurišić (Slovenia), Yaokun Wu (China), Mikhail Bounyaev (USA), Oleg Musin (USA), Sergey Shpectorov (UK), Fukshansky Lenny (California, USA), Navin Singhi (India), Marcelo Firer (Brasil), Yulia Kempner (Israel)

Редакционная коллегия:

доктор физико-математических наук, профессор В. Н. Чубариков;

доктор физико-математических наук, профессор Н. М. Добровольский;

кандидат физико-математических наук, доцент И. Ю. Реброва;

кандидат физико-математических наук Н. Н. Добровольский;

старший преподаватель А. В. Родионов

Алгебра, теория чисел, дискретная геометрия и многомасштабное моделирование: Современные проблемы, приложения и проблемы истории: Материалы XXIV Международной конференции, посвященной 110-летию со дня рождения академика Ю. В. Линника и 110-летию со дня рождения профессора А. Б. Шидловского и 80-летию со дня рождения профессора Г. И. Архипова.

Тула: Тул. гос. пед. ун-т им. Л. Н. Толстого, 2025. – 191 с.
ISBN 5–87954–388–9

ББК 22.1
УДК 51

ISBN 5–87954–388–9

© Тульский государственный
педагогический университет
им. Л. Н. Толстого, 2025

Эвристические алгоритмы учитывают не только бинарные отношения, связывающие действия между собой, но и частоту действий, частоту бинарных отношений и частоту экземпляров процесса. Результатом работы эвристических алгоритмов является С-сеть. Основная идея заключается в том, что редкие экземпляры процесса не следует включать в С-сеть. Тем самым эвристические алгоритмы теряют часть трасс исходного процесса и не могут строить точную С-сеть.

Генетические алгоритмы - это адаптивные методы поиска, имитирующие эволюционное развитие. Популяции эволюционируют путем отбора наиболее приспособленных особей и создания новых особей с использованием генетических механизмов, таких как скрещивание (объединение частей двух или более особей) и мутация (случайная модификация особи). Основой для выбора особей, реализации генетических механизмов и мутации является матрица причинно-следственных отношений.

Алгоритмы регионов основаны на понятии состояния процесса и переходе из одного состояния в другое при выполнении действия. Моделью такой системы является граф с множеством вершин, соответствующих состояниям, и дуг, соответствующим действиям. Такой граф называется системой переходов. Системы перехода просты, но имеют проблемы с точным выражением параллелизма. Используя некоторую функцию представления состояний, мы можем автоматически построить систему перехода на основе журнала событий L . Преобразовав журнал событий в систему переходов, мы можем синтезировать из неё сеть Петри. Задача состоит в том, чтобы объединить большую систему переходов в меньшую сеть Петри, обнаружив параллелизм. Основная идея заключается в обнаружении регионов, которые соответствуют позициям сети Петри. Регион - это набор состояний, при котором все действия в системе переходов "согласуются" с регионом.

СПИСОК ЦИТИРОВАННОЙ ЛИТЕРАТУРЫ

1. Wil M.P. van der Aalst. Process Mining. - Berlin: Springer, 2016. - 467 p.
2. Миронов А.М. Теория процессов.- Переславль-Залесский: Издательство НОУ Институт программных систем. 2008. - 346 с.
3. Котов В. Е. Сети Петри. — М.: Наука, 1984. — 160 с.
4. Alves De Medeiros, A. K., Dongen, van, B. F., Aalst, van der, W. M. P., Weijters, A. J. M. M. (2004). Process mining : extending the alpha-algorithm to mine short loops. (BETA publicatie : working papers; Vol. 113). Technische Universiteit Eindhoven.

УДК 519.16

Выравнивание последовательностей со структурой¹

Г. А. Хазиев (Россия, г. Москва)

Институт проблем передачи информации им. А. А. Харкевича РАН

e-mail: georgy.haziev@yandex.ru

А. В. Селиверстов (Россия, г. Москва)

Институт проблем передачи информации им. А. А. Харкевича РАН

e-mail: slvstv@iitp.ru

¹Работа выполнена в рамках государственного задания ИППИ РАН, утвержденного Минобрнауки России.

В. А. Любецкий (Россия, г. Москва)

Институт проблем передачи информации им. А. А. Харкевича РАН

e-mail: lyubetsk@iitp.ru

Structure-aware sequence alignment

G. A. Khaziev (Russia, Moscow)

Institute for Information Transmission Problems of the RAS (Kharkevich Institute)

e-mail: georgy.haziev@yandex.ru

A. V. Seliverstov (Russia, Moscow)

Institute for Information Transmission Problems of the RAS (Kharkevich Institute)

e-mail: slvstv@iitp.ru

V. A. Lyubetsky (Russia, Moscow)

Institute for Information Transmission Problems of the RAS (Kharkevich Institute)

e-mail: lyubetsk@iitp.ru

1. Введение

Вычисление (обобщённого) редакционного расстояния, поиск наибольшей общей подпоследовательности и выравнивание последовательностей широко используются в биоинформатике. Обобщённое редакционное расстояние, при котором замены и вставки дают разные вклады в расстояние, легко вычислить за квадратичное время [1]. Известно несколько алгоритмов для точного вычисления редакционного расстояния за меньшее время [2, 3]. Для фиксированной верхней границы на редакционное расстояние известны алгоритмы, время работы которых линейно зависит от длины последовательностей [4]. Программа MMseqs2 реализует быстрый эвристический алгоритм [5].

Для предсказания структуры белка уже создано несколько алгоритмов [6, 7]. Однако предсказание структуры по одной последовательности может быть трудным для белков с различными альтернативными структурами, формирование которых зависит от небольших изменений внешних условий. Другой причиной изменения структуры может быть модификация отдельных аминокислот.

Менее изучены задачи выравнивания последовательностей с учётом некоторой дополнительной структуры. Примером такой структуры, важной для изучения нуклеотидных последовательностей, служат несовершенные палиндромы, рассмотренные в недавней статье [8].

Мы рассмотрим аминокислотные последовательности со структурой, участвующей в образовании димеров или более сложных комплексов белков, состоящих из нескольких субъединиц. Такие структуры белков были впервые открыты в 1952 году Фрэнсисом Криком и названы coiled-coil. Известно много типов таких структур, образованных α -спиралями с повторяющимися гидрофобными аминокислотами [9]. Поскольку на один виток α -спирали приходится в среднем 3.6 аминокислот, два витка составляют близкое к целому число аминокислот. Поэтому для взаимодействия двух α -спиралей важно повторение гидрофобных аминокислот с периодом семь. Обычно рассматривают вырожденные повторы участка *abcdefg*, где на позициях *a* и *d* расположены гидрофобные аминокислоты, а на позициях *b*, *c* и *f* заряженные аминокислоты. Частный случай, когда позиции с периодом семь занимает лейцин, называется лейциновой молнией. Метод для предсказания в одной последовательности такого участка, который входит в состав структуры coiled-coil, был реализован в программе COILS, которую предложил Андрей Лупас (Andrei Lupas). Однако такое предсказание не учитывает выравнивания разных последовательностей.

Мы рассматриваем задачу выравнивания двух аминокислотных последовательностей, при котором вычисляется обобщённое редакционное расстояние и общие потенциальные участки, участвующие в образовании структуры coiled-coil.

2. Предварительные результаты

Метод Нидлмана–Вунша [1] состоит в последовательном заполнении таблицы (обобщённых) редакционных расстояний между всеми префиксами двух исходных последовательностей. В итоге в нижнем правом углу этой таблицы записано расстояние между последовательностями. Масек и Патерсон в 1980 году заметили, что этот метод требует меньше времени и меньшей памяти, если исходные последовательности разбить на блоки, которые могут повторяться. Таблица расстояний заполняется лишь для префиксов, состоящих из целого числа блоков.

3. Результаты

Задача многокритериальной оптимизации может быть сведена к задаче оптимизации одного функционала с заранее неизвестными параметрами, определяющими относительный вклад отклонения кадного из функционалов исходной задачи от оптимума.

Разметим в каждой из последовательностей (независимо от их выравнивания между собой) гептамеры — участки по семь аминокислот, соответствующие образцу для структур coiled-coil. Эти участки могут пересекаться между собой или быть разделены друг от друга любым числом аминокислот. Для всех пар гептамеров из разных последовательностей вычисляется расстояние при выравнивании без делеций. (Здесь видна аналогия с алгоритмом Масака–Патерсона.)

Рассмотрим наряду с основной таблицей расстояний между префиксами последовательностей, вспомогательные таблицы для других вариантов выравнивания.

Пусть известно оптимальное выравнивание (с оптимальной разметкой участков) для префиксов, длины которых меньше числа j для первой или числа k для второй последовательности. Для заполнения новых ячеек таблиц, рассмотрим несколько случаев.

Если хотя бы одна из позиций — j -я в первой или k -я во второй последовательности — не покрыта ранее размеченными гептамерами, то вычисляем расстояние между префиксами, длины которых равны j и k соответственно, как обычно в методе Нидлмана–Вунша.

Если обе позиции — j -я в первой и k -я во второй последовательности — покрыты ранее размеченными гептамерами, то параллельно продолжаем выравнивание двумя путями. На первом пути к префиксу добавляется гептамер, как это делалось в алгоритме Масака–Патерсона. Результаты вносятся во вспомогательную таблицу. На втором пути добавляем дополнительный штраф (за отказ от учёта гептамера) и продолжаем заполнение второй вспомогательной таблицы, пока не будет добавлено достаточно много (хотя бы семь) новых позиций. При этом может быть вторичное разветвление. Если на втором пути получено оптимальное расстояние для префиксов достаточно большой длины, то это значение сравнивается со значением на первом пути. Из этих вариантов выбирается меньшее. После того, как такой выбор сделан, значение расстояния записывается в основную таблицу и уже не меняется. Но из-за вторичных разветвлений окончательный выбор может быть отложен.

4. Заключение

Рассматриваемый нами метод может применяться для уточнения предсказания структуры белка и также для уточнения отношения гомологичности белков с учётом структуры. Малое время работы позволит проводить вычисления с большими наборами белков, в частности, при изучении эволюции эукариотических белков.

СПИСОК ЦИТИРОВАННОЙ ЛИТЕРАТУРЫ

1. Янковский В. О. Группы изометрий формальных языков относительно обобщённых метрик Левенштейна // Матем. заметки. 2024. Т. 116, № 2. С. 306–315.
2. Tiskin A. Semi-local longest common subsequences in subquadratic time // Journal of Discrete Algorithms. 2008. V. 6. No. 4. P. 570–581.

3. Grabowski S. New tabulation and sparse dynamic programming based techniques for sequence similarity problems // *Discrete Applied Mathematics*. 2016. V. 212. P. 96–103.
 4. Kociumaka T., Radoszewski J., Starikovskaya T. Longest common substring with approximately k mismatches // *Algorithmica*. 2019. V. 81. No. 6. P. 2633–2652.
 5. Steinegger M., Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets // *Nature Biotechnology*. 2017. V. 35. P. 1026–1028.
 6. Jumper J., Evans R., Pritzel A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021. V. 596. P. 583–589.
 7. Rachitskii P., Kruglov I., Finkelstein A.V., Oganov A.R. Protein structure prediction using the evolutionary algorithm USPEX // *Proteins*. 2023. V. 91. No. 7. P. 933–943.
 8. Зверков О. А., Селиверстов А. В., Шиловский Г. А. Выравнивание скрытого палиндрома // *Математическая биология и биоинформатика*. 2024. Т. 19, № 2, С. 427–438.
 9. Truebestein L., Leonard T. A. Coiled-coils: The long and short of it // *Bioessays*. 2016. V. 38. No. 9. P. 903–916.
-