МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования "Дальневосточный федеральный университет

Кафедра компьютерных систем

Разработка метода обработки больших геномных данных

Диплом на соискание степени бакалавра

Выполнил:
студент группы Б8117(09.03.02)
Делёва Элеонора Юрьевна

Научный руководитель:
(степень руководителя)
д.ф-м.н.
профессор Нефедев К.В.

Владивосток 2021

Содержание

1 РАЗРАБОТКА МЕТОДА ОБРАБОТКИ БОЛЬШИХ ГЕНОМНЫХ ДАННЫХ

1.1 Поколения секвенирования

Секвенирование ДНК - это технология, которая использует состав ДНК для понимания и расшифровки кода всей биологической жизни на Земле, а также для понимания и лечения генетических заболеваний [1]. Появление технологий секвенирования сыграло важную роль в анализе геномных последовательностей организмов. Секвенсор ДНК создает файлы, содержащие последовательности ДНК [2]. Эти последовательности представляют собой строки, называемые чтениями в алфавите, состоящем из пяти букв A, T, C, G, N. Обозначения представлены в таблице 1. Символ N используется для обозначения неоднозначности. Первые технологии секвенирования были разработаны в 1977 году Sanger et al. [3] из Кембриджского университета был удостоен Нобелевской премии по химии в 1980 г., а Maxam et al. [4] из Гарвардского университета. Их открытие положило начало изучению генетического кода живых существ и вдохновило исследователей на разработку более быстрых и эффективных технологий секвенирования. Секвенирование по Сэнгеру стало наиболее применяемым методом из-за его высокой эффективности и низкой радиоактивности [5] и было коммерциализировано и автоматизировано как «Технология секвенирования Сэнгера».

Каждый человек имеет свой собственный генетический код. Свойства такого кода представлены в таблице 2

Таблица 1: Номенклатура нуклеотидов для ДНК

Обозначение	Азотистое основание	Полное название нуклеотида
A	Аденин	Адениловый
G	Гуанин	Гуаниловый
C	Цитозин	Цитодиловый
T	Тимин	Тимидиловый

Таблица 2: Свойства генетического кода человека

Свойство	Описание свойства
Код триплетен	Кадой аминокислоте соответсвует сочетание 3-ёх нуклеотидов
Код обозначен	Каждый триплетный код соответствует только одной аминокислоте
Код вырожден	Каждая аминокислота имеет больше, чем один код
Код универсален	Все живые организмы имеют одинаковый код аминокислот
Код непрерывен	Между кодами нет промежутков

1.2 Результат секвенирования (пример)

Результат секвенирования ДНК по методу Сэнгера представлен на рисунке 1.

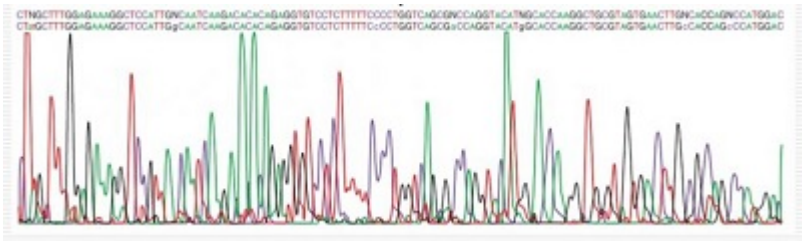

Рис. 1: Компьютерная запись флуорограммы одной из дорожек ПААГ геля и одного капилляра.

Технологии секвенирования Сэнгера и Максама-Гилберта были наиболее распространенными технологиями секвенирования, используемыми биологами до появления новой эры технологий секвенирования, открывающей новые перспективы для исследования и анализа геномов. Такие технологии секвенирования впервые появились в технологии 454 компании Roche в 2005 году [6] и были коммерциализированы как технологии, способные производить последовательности с очень высокой производительностью и гораздо более низкими затратами, чем первые технологии секвенирования. Эти новые технологии секвенирования широко известны под названием «Технологии секвенирования следующего поколения (NGS)» или «Технологии высокопроизводительного секвенирования».recomb

Для рассчета используется формула частоты рекомбинаций:

$$n_{rec} = \frac{N_p}{N_{ch}} * 100 \tag{1}$$

где N_p- общее число рекомбенантов, N_{ch} - общее число потомков.

1.3 Плавление ДНК

Вторичная структура ДНК играет важную роль в биологии, генетической диагностике и других методах молекулярной биологии и нанотехнологии.

Поэтому точное определение температуры плавления молекул ДНК или РНК играет самую главную роль во всех молекулярно-биологических методах (представлено на рисунке 2), например, при подборе проб или олигонуклеотидов для микрочипов или при подборе праймеров для ПЦР, особенно для секвенирования, изученого в параграфе 1.1 и 1.2.

Рис. 2: Процесс плавления ДНК

Состав нуклеотидов представлен на рисунке 3.

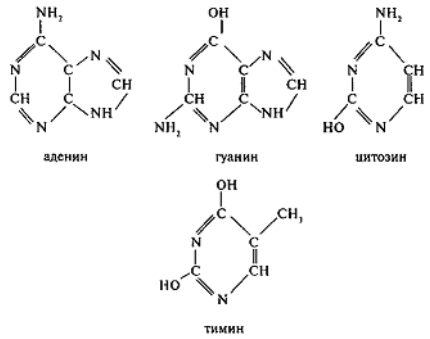

Рис. 3: Химический состав нуклеотидов

Существует несколько простых формул вычисления температуры плавления для коротких олигонуклеотидов. Грубое вычисление температуры плавления (Tm) короткого олигонуклеотида (<20 нуклеотидов) проводят по прямому подсчету количества нуклеотидов (G+C) — сумма всех гуанинов и цитозинов, L — длина олигонуклеотида):

$$T_m = 2(L + G + C) \tag{2}$$

Усредненная формула подсчета Tm для короткого олигонуклеотида (и для длинных фрагментов ДНК) с учётом концентрации ионов K+ и DMSO:

$$T_m = 77.1 + 11.7 \lg[K^+] + \frac{41(G + C) - 528}{L} - 0.75[\%DMSO] \tag{3}$$

Однако эти уравнения не учитывают инициацию связывания при гибридизации олигонуклеотида, не учитывают особенности самой последователь-

ности и концевого эффекта, характерный для олигонуклеотидных дуплексов. Поэтому данная формула пригодна в большей степени, где последовательность ДНК усредненная и длина дуплексов свыше 40 нуклеотидов. Для точного рассчета энергии взаимодействия между нуклеотидами и понимания, что происходит с белком в ДНК в работе используются формулы, рассмотренные выше (2) и (3).

Список литературы

[1] B. P. et al. Abbott. Observation of gravitational waves from a binary black hole merger. *Phys. Rev. Lett.*, 286, 2016.

[2] Hinsen K. Hugunin J. Dubois, P. F. Numerical python. *Comput. Phys.*, 1996.

[3] Dubois P. F. Hinsen K. Hugunin J. Oliphant T. E. Ascher, D. An open source project: Numerical python. 2001.

[4] A. et al. Chael. High-resolution linear polarimetric imaging for the event horizon telescope. *Astrophys. J.*

[5] T. E. Oliphant. Guide to numpy 1st edn. 2006.

[6] Furnish G. Dubois P. F. Yang, T.-Y. *Steering object-oriented scientific computations.* TOOLS USA 97, 1997.