Проблемно-ориентированый язык для быстрого поиска нуклеотидных последовательностей минимальной длины, удовлетворяющих различным топологиям связывания азотистых оснований

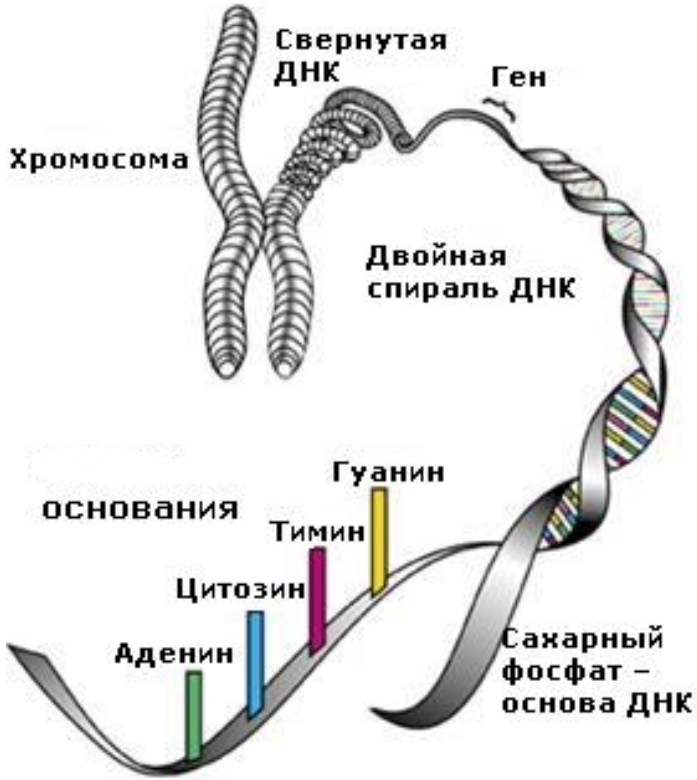
М. Юрушкин, Л. Гервич, С. Бачурин

Южный Федеральный Университет

5 апреля 2017 г.

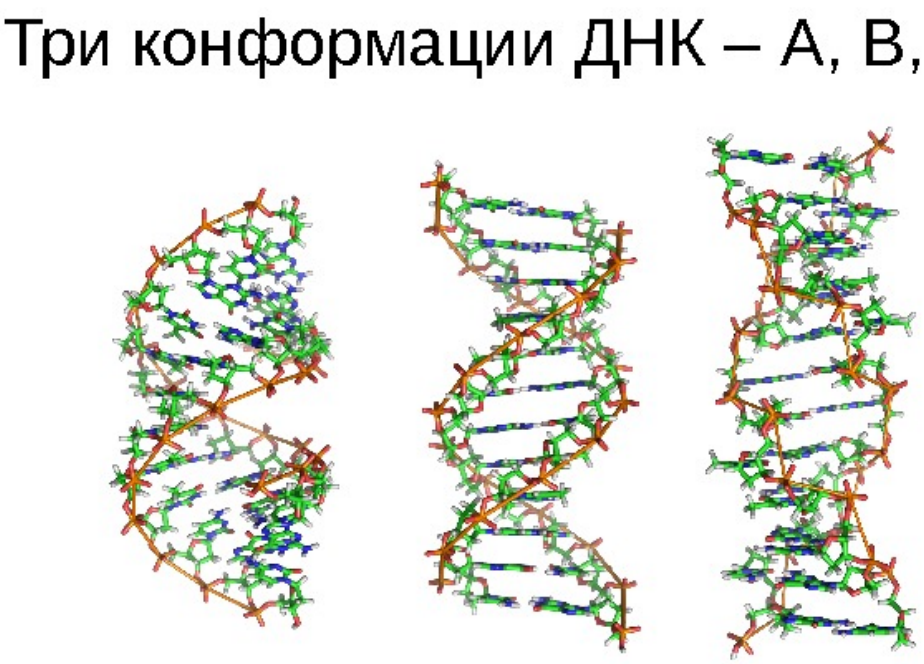
Несколько фактов из школьной биологии

- ДНК — полимерная молекула, содержащаяся в ядре клетки и отвечающая за наследственные признаки организма
- Структурная единица ДНК — **ген**. **Ген** состоит из нуклеотидной последовательности азотистых оснований: аденин, тимин, цитозин, гуанин



Искажения нативной формы ДНК

- Азотистые основания ДНК способны к образованию химических связей друг с другом и с иными веществами (интеркаляторами, металлами, протеинами)
- Это приводит к изменению общей формы спирали молекулы и переходу ДНК из нативной (B-форма) в менее упорядоченные (A,C,Z и т.д. формы)



Неканонические структуры ДНК

Так называемые неканонические структуры ДНК, имеющие следующие нуклеотидные последовательности, вносят основной вклад в дестабилизацию нативной спирали ДНК:

- G-квадруплекс: $X_a G_m X_b G_m X_c G_m X_d G_m X_e$
- I-мотив: $X_a C_m X_b C_m X_c C_m X_d C_m X_e$
- Шпилька: $(ATGC)_a X_b (GCAT)_a$
- Триплекс: $(Pyr)_a X_b (Pur)_a X_c (Pyr|Pur)_a$

где G – гуанин, C– цитозин, A – аденин, T – тимин, Pyr – цитозин или тимин, Pur – гуанин или аденин

Цель и задачи исследования

Цель исследования: найти нуклеотидную последовательность минимальной длины, способную образовывать все виды неканонических структур ДНК. Предполагается, что такие участки ДНК наиболее лабильны к воздействиям, искажающим B-форму ДНК.

Задачи исследования:

- Описать неканонические структуры ДНК
- Разработать алгоритм поиска последовательности минимальной длины, удовлетворяющей всем неканоническим структурам

Проблема описания неканонических структур ДНК

В биоинформатике возникает задача поиска мотивов в последовательности. Ее решают с помощью регулярных выражений. Здесь задача оказалась совершенно другой.

- Уже используемый в биоинформатике аппарат регулярных выражений, позволяющий решать задачи поиска мотивов в геномной последовательности, нам не подходит
- Как оказалось, регулярные выражения не позволяют емко описать неканонические структуры ДНК

Проблема описания неканонических структур ДНК

Требуется:

- возможность компактного описания неканонических структур языка
- возможность дальнейшего переиспользования и расширения описания

Нужен гибкий язык описания неканонических структур, а также операций над ними.

Грамматики неканонических структур

- $GQD = X^*g\{m\}X\{3\}X^*g\{m\}X\{3\}X^*g\{m\}X\{3\}X^*g\{m\}X^*, m = [1:20]$
- $IMT = X^*c\{m\}X\{3\}X^*c\{m\}X\{6\}X^*c\{m\}X\{3\}X^*c\{m\}X^*, m = [1:20]$
- $HRP_1 = X^*a\{m\}t\{n\}c\{p\}g\{r\}X\{4\}X^*c\{r\}g\{p\}a\{n\}t\{m\}X^*, m = [1:5], n = [1:5], p = [1:5], r = [1:5]$

Грамматики неканонических структур

- $TRP_1 = X^*aX\{4\}X^*tX\{3\}X^*gX^*$
- $TRP_2 = X^*cX\{4\}X^*gX\{3\}X^*tX^*$
- $TRP_3 = X^*tX\{4\}X^*aX\{3\}X^*gX^*$
- $TRP_4 = X^*cX\{4\}X^*aX\{3\}X^*tX^*$
- ...
- $X = (a|c|g|t)$
- $result =$
  $(GQD)|(IMT)|\,(HRP_1)|\,(TRP_1|TRP_2|TRP_3|...|TRP_27|TRP_28)$

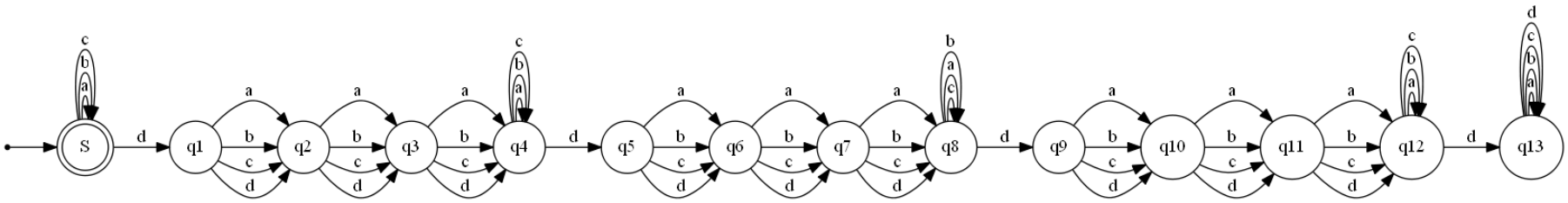
Алгоритм нахождения минимальной строки, удовлетворяющей топологиям

1. На первом шаге каждый используемый паттерн, заданый во входной программе, конвертируется в NFA. Затем каждый полученный недетерменированный автомат конвертируется в DFA
2. Строится граф вычислений, в котором вершинами являются операции над автоматами (объединение, пересечение и т.д.).

Алгоритм нахождения минимальной строки, удовлетворяющей топологиям

3. Над каждым полученным DFA производится минимизация с помощью алгоритма Мура
4. В результате выполнения всех операций в графе вычислений получается результирующий DFA. Все строки, которые полученный DFA допускает, удовлетворяют заданным топологиям. Полученный DFA рассматривается как ориентированный граф $G(V, E)$. В графе $G(V, E)$ осуществляется поиск минимального пути с помощью алгоритма Дейкстры из вершины, соответствующей начальному состоянию DFA

Пример DFA для G-квадруплекса



Результаты

Некоторые минимальные полученные строки

CGATCGCATCTCGATCG
CGATCGGATCTCGATCG
CGATCGATTCTCGATCG
CGATCGTTTCTCGATCG
CGATCGCTTCTCGATCG
CGATCGGTTCTCGATCG
CGATCGACTCTCGATCG
CGATCGTCTCTCGATCG
CGATCGCCTCTCGATCG