Методы аннотации геномов квадруплексами, Z-ДНК, структурами стебель-петля.

1. Квадруплексы

Классические методы: Спектроскопия кругового дихроизма
Экспериментальные высокопроизводительные методы:

- G4-seq
- G4-ChIP

Published online 2005 May 24. doi: 10.1093/nar/gki609

Prevalence of quadruplexes in the human genome

Julian L. Huppert and Shankar Balasubramanian*

Предпосылки:

1. Стабильность увеличивается с увеличением числа тетрад (стэкинг)
2. Стабильность уменьшается с увеличением длины петель.
3. Были обнаружены структуры с петлями от 1 до 7
4. Одиночные тетрады были обнаружены только при высоких концентрациях гуанина, физиологическое значение маловероятно
5. Квадруплексы из двух тетрад имеют физиологическое значение, но стабильность очень низкая

Предложено правило:

‘следовательность типа $d(G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+})$ сворачивается в квадруплекс при условиях, близких к физиологическим.’

Реализация:

- quadparser
- регулярные выражения

1.1 Поиск по паттерну

$d(G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+})$

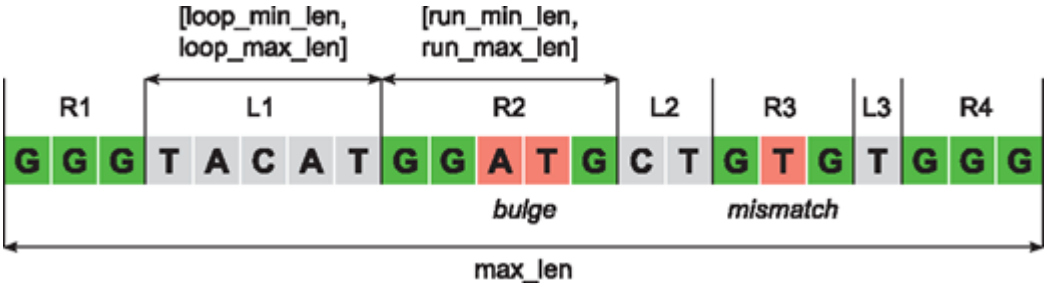
Вспомним основной синтаксис регулярных выражений

()- группа

{m,n}-повторение от m до n раз
[]- символьный класс или набор символов. У нас ДНК, поэтому [ATGC]. (Хотя в данном случае, сойдет и .)
(G{3,}[ATGC]{1,3}){3,}G{3,}
Что делать с пересекающимися?
GGGGAGTCGGGGAAAGGGGATCTGACGGGG
GGGGAGTCGGGGAAAGGGGATCTGACGGGG

GGGGAGTCGGGGAAAGGGGATCTGACGGGG
GGGGAGTCGGGGAAAGGGGATCTGACGGGG
Вводить score
Как его вводить?
Пример

pqsfinder https://academic.oup.com/bioinformatics/article/33/21/3373/3923794



$$S_r = (N_t - 1)B_t - N_m P_m - \sum_{i=1}^{N_b} P_b + F_b L_{bi}^{E_b}$$

N_t the number of tetrads,
B_t a G-tetrad stacking bonus,
N_m the number of inner mismatches,
P_m mismatch penalization,
N_b the number of bulges,
P_b bulge penalization,
F_b bulge length penalization factor,
L_{bi} the length of the i-th bulge
E_b bulge length exponent.

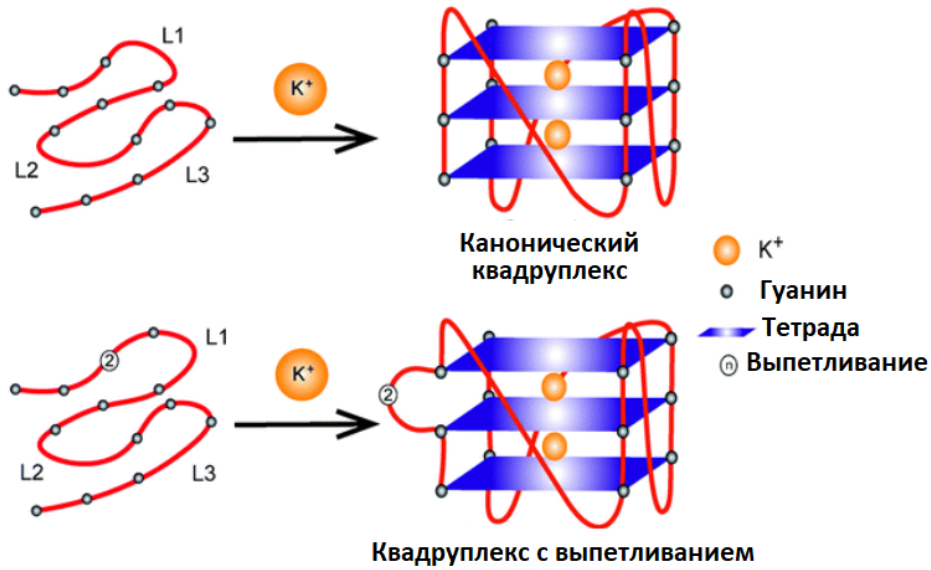
$S=\max(S_r - F_m L_{Emm}, 0)$

F_m , E_m numerical parameters that empirically model the relationship between loop lengths and their destabilization effects on the quadruplex

Практика

```
import re
#устанавливаем biopython если нет командой pip install biopython
from Bio import SeqIO
#скачиваем архив с 22-й хромосомой hg19 и разархивируем
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/chr22.fa.gz
input_file = "C:\\Users\\...\\Downloads\\chr22.fa"
fasta_sequence = SeqIO.parse(input_file,'fasta')
for record in SeqIO.parse(input_file, "fasta"):
    print("%s %i" % (record.id, len(record)))
    name, sequence = record.id, str(record.seq)
pattern="(?:G{3,}[ATGC]{1,7}){3,}G{3,}"
PQS=[[m.start(),m.end(),m.group(0)] for m in re.finditer(pattern,sequence)]
len(PQS)
#Должно получиться 3542
```

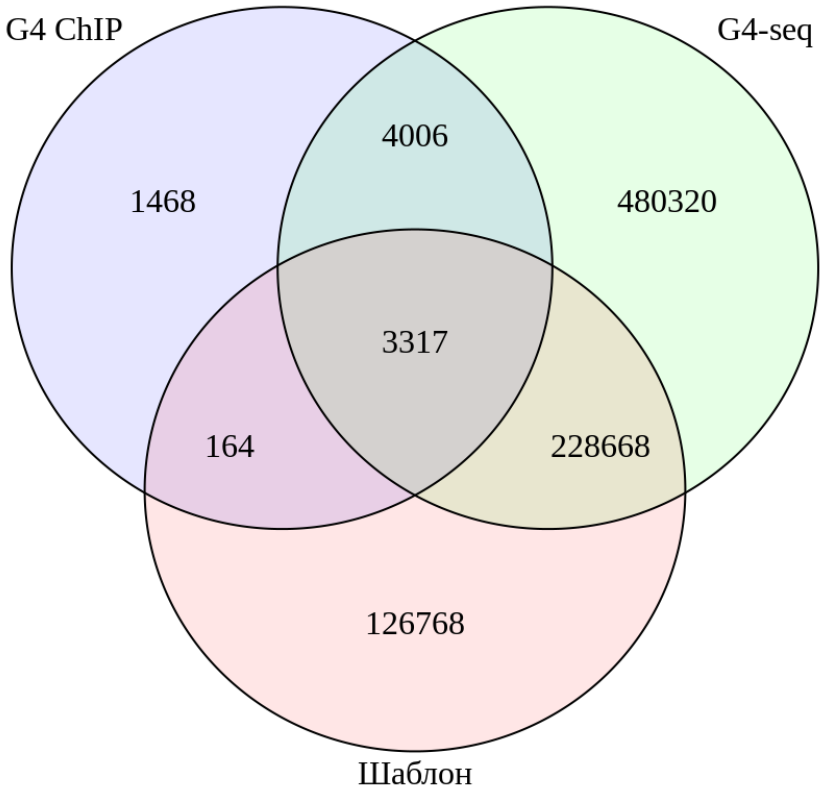
Для сравнения, скачиваем данные ChIP-seq
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107690
```
import pandas as pd
chip=pd.read_csv("C:\\....\\Downloads\\GSE107690_K562_High_confidence_peaks.bed\\GSE107690_K562_High_confidence_peaks.bed", sep="\t",names=['chr', 'start', 'end'])
chip.head()
ip[chip.chr=="chr22"].shape
```
#(251,3) Получили 251



По паттерну 3542, а в ChIP-seq 251. Почему?

- выпетливания
- длинные центральные петли
- квадруплексы из двух тетрад
- нуклеосомы

На самом деле, всё ещё хуже.



Extra*:
G4-Hunter http://bioinformatics.ibp.cz:8888/#/analyse/quadruplex
Quadron http://quadron.atgcdynamics.org/

Посмотреть таблицу из обзорной статьи по методам
И другие https://academic.oup.com/view-large/185803057

2. Z-DNA

Полезные ссылки

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1167176/

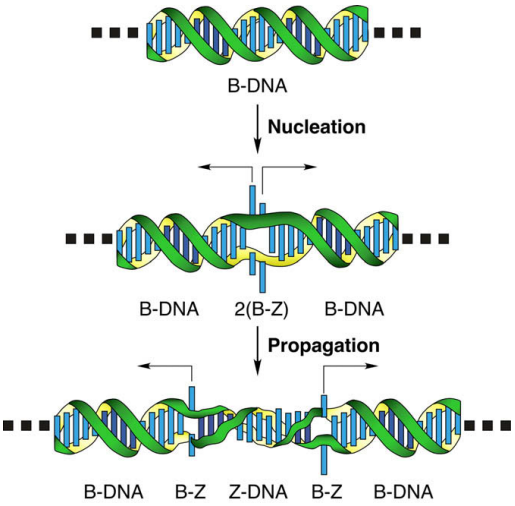
https://doi.org/10.1016/j.ymeth.2008.09.007

Z-hunt and Z-catcher больше не работают!

НЕ РАБОТАЕТ https://wiki.christophchamp.com/index.php?title=Z-Hunt
НЕ РАБОТАЕТ
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.685.244&rep=rep1&type=pdf
http://vhp.ntu.edu.sg/zdna/Z_Catcher.zip

РАБОТАЕТ:
https://nonb-abcc.ncifcrf.gov/apps/nBMST/default/, с критерием поиска "G followed by Y (C or T) for at least 10 nt; One strand must be alternating Gs"

Теория
Z-hunt - термодинамический подход



$$S_j = e^{-\left(\Delta G^o_{B-Z} + K[\Delta Lk - j\{-0.36\text{turns/dn}\}]^2\right)/RT}$$

S -propagation terms
ΔG_{B-Z} - the propagation free energy for a given dinucleotide,
j - is the number of dinucleotides propagated
0.36 - is the change in twist associated with the formation of Z-DNA in a dinucleotide

$$\Delta G^\circ = K(\Delta Lk_m - \Delta Tw_m)^2$$

$$\sigma = e^{-(10\ \text{kcal/mol} + K[\Delta Lk - (-0.8)]^2)/RT}$$

σ - nucleation term
0.8 - the unwinding associated with the formation of the B–Z junctions.

$$Q = 1 + \sum_{i=1}^{n} \sum_{k=1}^{n} \sigma\left(\prod_{j=i}^{k} S_j\right) \exp\left\{\frac{-K}{RT}\left(\Delta Lk - \left[\sum_{j=i}^{k} 0.36j\right] - 0.8\right)^2\right\}$$

$$< \Delta Tw > = Q^{-1}\left[\sum_{i=1}^{n} \sum_{k=1}^{n} \left(\left[\sum_{j=i}^{k} 0.36j\right] - 0.8\right)\sigma\left(\prod_{j=i}^{k} S_j\right) \times \exp\left\{\frac{-K}{RT}\left(\Delta Lk - \left[\sum_{j=i}^{k} 0.36j\right] - 0.8\right)^2\right\}\right]$$

partition function (Q)
n number of dinucleotides
Tw - helical twist
0.179 - ΔTw for each dinucleotide
0.4 - ΔTw of the two B–Z junctions.
The amount of Z-DNA can thus be predicted as the change in overall twist of the plasmid at any ΔLk
ΔWr superhelical writhe
ΔLk change in linking number

pattern="([AG][CT]){10,}"

3. Шпильки - Hairpins

Программа DNA punctuation
https://github.com/mariapoptsova/dnapunctuation

алгоритм поиска (метод динамического програмирования) описан в Методах тут:
https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3344-4

Посмотреть файл README.md

1. Скомпилировать программу
>g++ dnapunctuation.cpp -o dnapunctuation

2. Запуск с разными параметрами

(A)
StemMin=6
StemMax=15
LoopMin=3
LoopMax=10
Mismatches=1
Последовательность транспозона L1 лежит на гитхабе

>dnapuncutation L1.fna 6 15 3 10 1

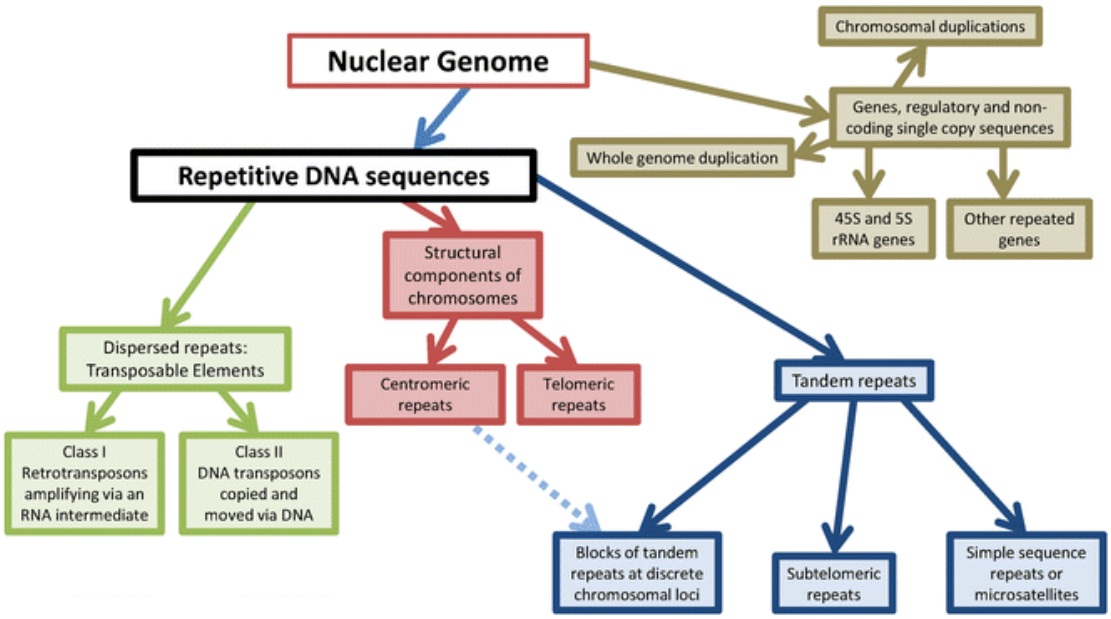
(B) StemMin=15
StemMax=30
LoopMin=3
LoopMax=10
Mismatches=3
>dnapuncutation L1.fna 15 30 3 10 3

3. Посмотреть результаты в соответствующем файле L1.fna.S6-15_L3-10_M1.pal

Если останется время, можно начать разбирать повторы.

4. Repeats finder

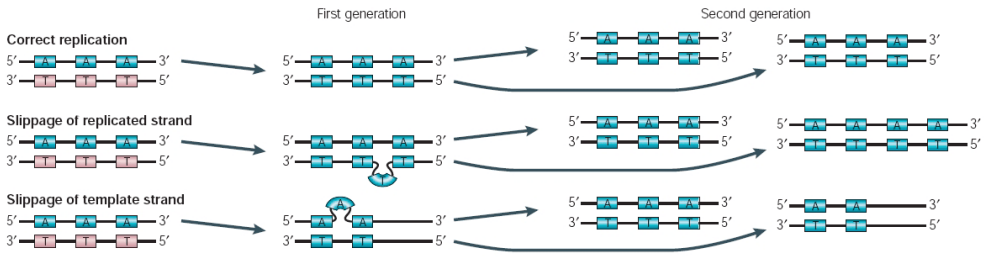
4.0 Виды повторов



Источник: Biscotti, Maria Assunta, Ettore Olmo, and JS Pat Heslop-Harrison. "Repetitive DNA in eukaryotic genomes." (2015): 415-420.

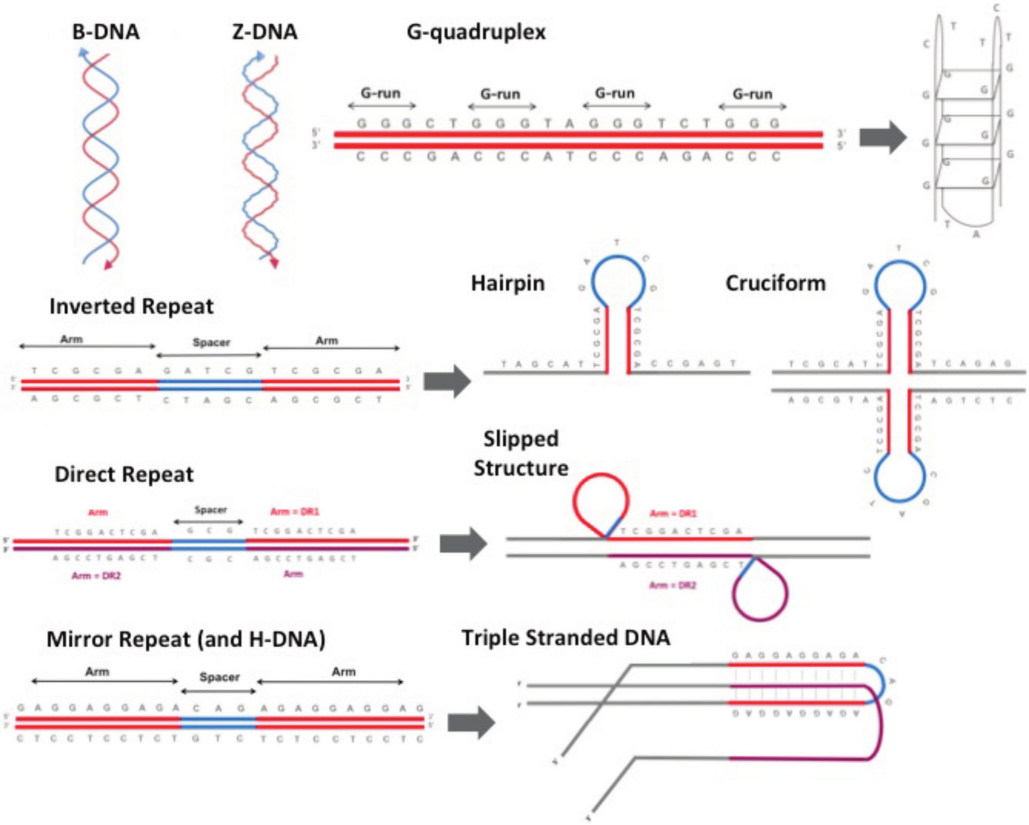
Различают тандемные (примыкающие друг к другу) и диспергированные (распределенные по геному) повторы.

Тандемные повторы: сателлитные, минисателлитные (10-60 п.н.), микросателлитные (1-6 п.н.). Участки ДНК с такими повторами отличается повышенным GC-составом, встречаются в эукариотов в области теломер, центромер, придают повышенную “хрупкость”. Вызывают ошибки репликации при редактировании (*slipped strand mispairing*) -> вставка или удаление повторов.



Источник: Myers P. (2007). Tandem repeats and morphological variation. *Nature Education*. **1**, 1;

Вторичные структуры ДНК



Источник: Georgakopoulos-Soares, Ilias, et al. "Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis." *Genome research* 28.9 (2018): 1264-1271.

Транспозоны и ретротранспозоны

ДНК-транспозоны: “вырезать и вставить”, на конце инвертированные повторы, распознаваемые транспозазой.

Ретротранспозоны: “копировать и вставить”, длинные концевые повторы, короткие диспергирующие повторы и тд.

4.1 RepeatMasker

http://www.repeatmasker.org/

56% of human genomic sequence is annotated by repeats

Программа ищет:

- Simple Repeats - Duplications of simple sets of DNA bases (typically 1-5bp) such as A, CA, CGG etc.
- Tandem Repeats - Typically found at the centromeres and telomeres of chromosomes these are duplications of more complex 100-200 base sequences.

- Segmental Duplications - Large blocks of 10-300 kilobases which are that have been copied to another region of the genome.
- Processed Pseudogenes, Retrotranscripts, SINES - Non-functional copies of RNA genes which have been reintegrated into the genome with the assitance of a reverse transcriptase.
- DNA Transposons
- Retrovirus Retrotransposons
- Non-Retrovirus Retrotransposons (LINES)

Скачиваем последовательность длиной 40kb 21 хромосомы из Table browser UCSC.
(последовательности > 50kb необходимо запускать на локальной версии программы)
Идём в Services: RepeatMasking.
Запускаем (файл большой -> указываем email)
Пример результата запуска:
http://www.repeatmasker.org/tmp/c96442ccd187454edaac4a6692831c36.html
Там же посмотреть ссылки на masked file и alignment.

Аннотация встроена в UCSC Genome Browser, можно посмотреть её прямо в браузере.

4.2 Non-B Database (https://nonb-abcc.ncifcrf.gov/apps/site/default)
База данных последовательностей ДНК, способных формировать вторичные структуры.

Данные в bed-формате доступны по адресу: https://isp.ncifcrf.gov/isp/nonb_dwnld/
Посмотреть визуализацию и статистику на сайте.

Помимо самих данных, есть tool для поиска такого рода мотивов в любых последовательностях - non-B DNA Motif Search Tool.
Запустить ту же последовательность, сравнить результаты с Repeat Masker.
статья: https://www.ncbi.nlm.nih.gov/pubmed/22470144?dopt=Abstract

4.3 Tandem Repeats Finder (extra, если останется время)
http://tandem.bu.edu/trf/trf.html

Списки других программ для поиска повторов (на деле их больше):
https://molbiol-tools.ca/Repeats_secondary_structure_Tm.htm
https://bioinformaticsonline.com/pages/view/27459/tools-for-searching-repeats-and-palindromic-sequences