Е. С. Черемушкин

АНАЛИЗ РАЗЛИЧНЫХ УЧАСТКОВ ДНК С ПОМОЩЬЮ АВТОКОРРЕЛЯЦИОННОЙ ФУНКЦИИ[1]

ВВЕДЕНИЕ

Функционально различные участки ДНК имеют различную структуру. В настоящее время общие закономерности различных функциональных участков ДНК подробно изучаются. В данной статье проведен анализ таких участков ДНК с помощью автокорреляционной функции и выявлены некоторые интересные особенности структуры различных функциональных участков.

Обычно целью данного изучения ставится распознавание неизвестных участков ДНК. Нашей целью было изучение структуры ДНК: найти некоторые различия, которые возможно не будут способствовать распознаванию неизвестных участков, но которые характеризуют качественные различия ДНК разных функций.

Первый этап такого исследования — это изучение автокорреляционной функции сигнала, образованного последовательностью ДНК.

1. СРАВНЕНИЕ АКФ РАЗЛИЧНЫХ ТИПОВ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Для распознавания неизвестных участков в ДНК наилучшими оказались методы скрытых марковских моделей, методы нейросетей и частотный анализ [1]. Но эти методы не дают достаточного представления об особенностях сигналов в этих участках. Теряются даже очень простые, но в то же время интересные закономерности.

Для первого эксперимента были выбраны три типа последовательностей: регуляторные районы (районы, с которыми связываются специфические белки — транскрипционные факторы — и которые отвечают за регу-

[1] cher@bionet.nsc.ru

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 05-08-33448).

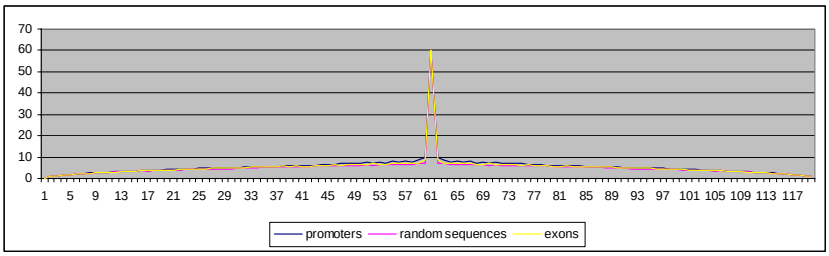
ляцию процессов в клетке), случайные последовательности и кодирующие районы (районы, которые кодируют белок).

Преобразуем ДНК в комплексную определяющую последовательность согласно таблице 1. Разобьем участки на районы длины 60. Теперь построим для каждого района $s_i = a_1, ..., a_N$ АКФ по формуле $F_i(t) = |sum(a_k * a_{k+t}^)|/L$, где ^ — комплексное сопряжение, L — длина последовательности, в данном случае равная 60. Усредним $F_i(t)$ по всем районам данного типа. Полученная средняя АКФ, представлена на рис. 1.

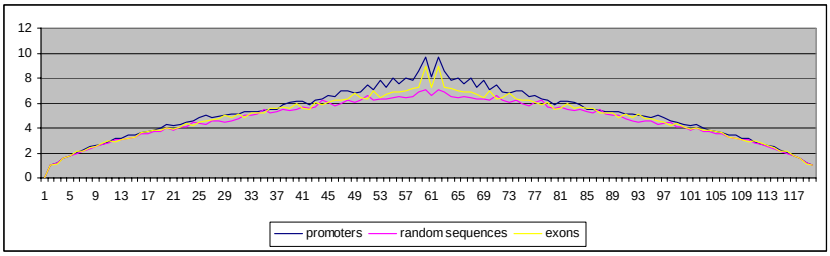
Таблица 1

Преобразование последовательности ДНК в определяющую последовательность сигнала

A	(1,0)	G	(0,-1)
C	(0,1)	T	(-1,0)



а)



б)

Рис. 1. а) АКФ различных районов ДНК; б) АКФ без нулевого пика

Видно, что АКФ различных участков практически совпадают, но видимые различия статистически значимы. Средняя корреляция этих районов: регуляторные районы (промоторы) — 5.007, случайные последовательности — 4.543, кодирующие районы (экзоны) — 4.707. Таким образом, видим, что регуляторные районы более скоррелированны, чем экзоны, а те более скоррелированны, чем случайные последовательности. Для того чтобы проверить достоверность этих данных, получим аналогичные значения для половины выборок: регуляторные районы (промоторы) — 4.988, случайные последовательности — 4.529, кодирующие районы (экзоны) — 4.765. Таким образом, можно утверждать, что данные о различиях в корреляции достоверны.

2. АНАЛИЗ РАЗБРОСА АКФ В РАЗЛИЧНЫХ УЧАСТКАХ ДНК

Рассмотрим дисперсию АКФ различных районов ДНК. Дисперсия характеризует разнообразие участков ДНК внутри одного функционального класса. Если АКФ внутри участков одного типа сильно варьирует, то это означает функциональную насыщенность фрагментов данного типа. Природе «необходимо», чтобы некоторые участки были крайне уникальны, а другие наоборот — очень похожи.

Посчитаем среднее значение E_i боковых пиков АКФ всех последовательностей данного типа. Затем посчитаем дисперсию по выборке $\{E_i\}$. Получим следующие значения: регуляторные районы (промоторы) — 26.03, случайные последовательности — 8.12, кодирующие районы (экзоны) — 11.66. Таким образом, промоторы гораздо более разнообразны по своей информационной насыщенности, чем другие последовательности. Это объясняется их назначением: они содержат участки, которые должны распознаваться специфическими белками. Причем определенные белки должны связываться только с конкретными промоторами. Если более подробно взглянуть на участки с маленькой и большой АКФ, можно заметить их отличие от случайных последовательностей (рис. 3).

Низкая АКФ

```
aactgaggtgcacagaggaaacctagttaagtactaagtggtggacttagaattttgaag
aatggttgtaagccaccaggtggttgctgggatttgaactccgggacctctggaagagca
agagaactctatataaatcaggttatttggcaggggtccaagatataacccaagccgaa
ccccggttatctaccggggctgggggttggtgttaaaattctcaagctagtatgtgctca
ccgttttctagggcaggaacccagggagggaagggctctggctggtatactgtaagtagc
cggtgttttagtgggtacatttcaacacaagcagcatttgggattacatcagcaaataca
cgtctggaggcttggttctgaagatgtggttttatcctcagctcagctcaaaattcgctg
ctaacgttaggctggaggtttcccgatagagaattcccctcacaccctctgccaatcctc
ctgggaaccaaacttgtatcttatgaaatagtagtaaagctctcattacagttacagctg
gcaggaacgtgcgcaccgccggagcctggcggcgacagtcacctgctcagcgggtacaga
gcgcagggaccacgctccctacccgccccgctggcagccgcagacagggcgccccgagaa
ggtcttggttttaggtcttttcaaagactaagcaatcttgttccgagctagcttttgagg
taacctgctgaaatccgaggtcgggctgttgctaggatgggcacctctccctctagtgtt
tatagacagcgaagagcccctgcagccgagctaggccagaagaacccatggctggaactc
ttaaactttgtgtgcctcagtttccttatctgtaaagtgggattaataatgtcatcctga
tttaaaattacagttgacctgggacgaagttcccataaagacctgcaagtcatccaccat
```

Высокая АКФ

```
aaaaaaaaaaaaaaaaagttcaaattctaactctgtaacagaaaccgcgaaaaaaaaccca
aaaaggaggtgggctatttctgtttatccacctgtctctacaaaagcaaataaaactgtg
acacacacacacacacacacaggctcttagtgcatgtatccacctctctctctctctc
acagagaaaccctgtctcaaacaaacaaacaaacaaacaaacaaacaaacaaacaacacc
atcacctttaatgtatttgcattttattatatacacatatacaaacatatatacatacat
cccagctctccatagccctgccccagctccccacagccctgccccagctccccagaccca
ccccagttccccacagccctgccccagctctccatagccctgccccagctctccatagcc
ctctttctttctttctttctttctctctctctttccttctttctttctttctttct
gagaaagaagtgagaggagggaagacaggggaagggaagagagaaaagagaggagagttg
gcccgcgcccgagcccaggccgagccttgccctggtccccggccgggccgaggccgcgcc
ggggcggtgcgggggggggggggaagggagtggtctccataagggggaggggagaaggcag
gtgcctagctctccgttgttgcaggggggaaaaaaaagtattgggaaaaaaaaaaaaccat
tacaagagaaaaaaaacaaacaaacaagaaacagctacaacccggcaaacgaaccacgtg
tatgttttcttttgctctctttctccctttctccctttctttctttctctttttctttct
tcgcgctccgcccttctccgtcctcttagcctcgctccttccctcagccccgccctcaat
tctctcacacacacacacacacacacacacacacacacacacaccacaaatgcattggtgtt
tctctctctttctttctctttctcatatgcacacatgcacacacacacataactctgaac
ttccttttctttttctttctctttttttccttttccaaaaggggtcacacacacacccgcc
ttctttctttctttctttctttctttctttctttctccttttttttgtttgtttggctttt
tttacattttgttatatacatacataatacacacatacacatatatacatacatataact
```

Рис. 3. Участки промоторов с низкой и высокой АКФ

3. ИЗУЧЕНИЕ ЗАВИСИМОСТИ СКОРРЕЛИРОВАННОСТИ РЕГУЛЯТОРНЫХ УЧАСТКОВ ОТ ПОЛОЖЕНИЯ В ГЕНЕ

Для предварительного анализа были взяты промоторные участки ДНК в районе старта транскрипции от −1000 до 80. В данном эксперименте участки были разбиты на 3 района: R1 = [−1000, −640], R2 = [−640, −280], R3 = [−280, 80]. Значения среднего и дисперсии в этих районах A(R1) = 5.04, A(R2) = 5.10, A(R3) = 4.86, D(R1) = 18.34, D(R2) = 15.52 и D(R3) = 9.88.

Таким образом, ближе к старту транскрипции средняя АКФ у промоторов понижается, и это говорит о том, что в районе старта больше достаточно уникальных участков. Из этого можно сделать вывод: низкая АКФ участка указывает на то, что этот участок несет функциональную нагрузку. Из этого следует, что природе «выгодно» поддерживать общее однообразие ДНК, и только функционально важные участки имеют свой «уникальный» паттерн.

ЗАКЛЮЧЕНИЕ

Суммируя результаты исследований, можем заключить, что АКФ различных участков практически совпадают, но видимые различия статистически значимы. А именно, регуляторные районы более скоррелированны, чем экзоны (кодирующие районы), а те более скоррелированны, чем случайные последовательности.

Промоторы (регуляторные участки) — гораздо более разнообразны по своей информационной насыщенности, чем другие последовательности. Это объясняется их назначением: они содержат участки, которые должны распознаваться специфическими белками. Причем определенные белки должны связываться только с промоторами одного типа.

Ближе к старту транскрипции средняя АКФ у промоторов понижается, и это говорит о том, что в районе старта больше достаточно уникальных участков. Из этого можно сделать вывод, что низкая АКФ участка указывает на то, что этот участок несет функциональную нагрузку. Таким образом, природе «выгодно» поддерживать общее однообразие ДНК и только функционально важные участки имеют свой «уникальный» паттерн.

СПИСОК ЛИТЕРАТУРЫ

1. **Azad R.K., Borodovsky M.** Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory // Brief Bioinform. — 2004. — Vol. JU-5, N 2. — P. 118–130.