ОПРЕДЕЛЕНИЕ ИНФОРМАЦИОННОЙ ЕМКОСТИ НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

УДК 575.21

Н.Н. Бугаенко[1], А.Н. Горбань[1], М.Г. Садовский[2]

[1]Вычислительный Центр СО РАН, Красноярск

[2]Институт биофизики СО РАН

Рассматривается проблема определения информационной емкости нуклеотидных последовательностей. Получены выражения для восстановления частотных словарей высших порядков по низшим. Описаны особенности информационных характеристик реальных нуклеотидных последовательностей, достоверно отличающие их от случайных текстов.

Ключевые слова: информация, энтропия, генетический текст.

1.Введение

В настоящее время идет быстрое накопление данных о нуклеотидных последовательностях организмов различных видов. Все возрастающий объем такой информации требует специальных методов и подходов для ее осмысления.

При информационном анализе генетических текстов возникает ряд проблем [1-5]. Прежде всего, это проблема формального определения объектов анализа. Затем - проблема сравнения информационного содержания текстов различной длины. Интересна также и проблема выделения информационных подструктур заданного текста.

2. Формальное описание объекта исследования

2.1. Частотные словари

Предметом нашего исследования является количественное определение наследственной информации, содержащейся в участках молекул ДНК либо РНК. Нуклеотидные последовательности мы рассматриваем как текст, т.е. как последовательность символов из четырехбуквенного алфавита **A**,**C**,**G**,**T**. *Длина* генетического текста - количество N нуклеотидов в нуклеотидной последовательности.

Предполагается, что клеточные механизмы, связанные со считыванием наследственной информации, оперируют с отдельными, весьма малыми, фрагментами НК. В соответствии с этим перейдем от рассмотрения молекулы как целого к изучению совокупности ее фрагментов фиксированной длины (*слов*). Список всех слов длины q, входящих в данный текст, называется *q-носителем* данного текста. Если каждому слову q-носителя сопоставить частоту его встречаемости в тексте, получим *частотный словарь длины q* [8].

Переход от текста к его частотному словарю - простой, но продуктивный прием, позволяющий единообразно работать с текстами различной длины, сравнивать их, производить информационный анализ. Кроме того, частотный словарь фиксирует информацию о тексте в совокупности небольших объектов - слов с их частотами, которые могут храниться по отдельности, "россыпью".

2.2.Задача восстановления словарей

Из словаря всегда можно путем суммирования получить словарь меньшей длины. При этом часть информации теряется: в общем случае обратное восстановление (от меньшей длины к большей) невозможно. Базовой в исследовании частотных словарей является *задача восстановления* (текста - по словарю, большего словаря - по меньшему).

Зададимся следующими вопросами:

- какую часть информации о тексте содержат его словари разных длин?
- какой из двух словарей содержит больше информации?
- как отличаются по информативности словари случайных и реальных генетических текстов?

Известно [7, 8], что начиная с некоторой длины d^* текст полностью восстанавливается по своим словарям длины $>d^*$. Эта длина для генов составляет, как правило, от 10 до 20 нуклеотидов. Характерные значения $d^*/\log_2 N$ для генов (мРНК) и вирусов человека (мРНК) лежат в интервале 1,2 $\pm$ 0,05. Нормировка на $\log_2 N$ обусловлена тем, что для случайных текстов величина $d^*/\log_2 N$ не зависит от N.

Словарь длины $q>d^*$ содержит всю информацию о тексте, поэтому поставленные вопросы содержательны только для словарей длины $q<d^*$. Для случайного текста длины N и натурального $l<N$ вероятность того, что $d^*>l$, допускает простую оценку:

$$\frac{1}{2}N^2(f_1^2 + f_2^2 + \ldots + f_k^2)^l$$

где k - число букв в алфавите, f_i - частота (вероятность), с которой i-я буква встречается в тексте, и предполагается, что появление букв на различных местах текста происходит независимо.

3.Метод максимума энтропии и восстановление словарей

Будем искать такое продолжение данного словаря, которое не привносит в восстанавливаемый словарь дополнительной информации, и следовательно обладает наименьшей определенностью. Этот словарь-продолжение должен обладать наибольшей энтропией, традиционно определяемой как

$$S_q = -\sum_{i_1 \ldots i_q} f_{i_1 \ldots i_q} \ln f_{i_1 \ldots i_q}, \quad (1)$$

где k -длина словаря, $i_1 \ldots i_q$ - слово из носителя словаря, $f_{i_1 \ldots i_q}$ - частота встречаемости этого слова, а суммирование производится по всем словам носителя. В дальнейшем буквы *i, j, k, ...*, с индексами или без, обозначают отдельные нуклеотиды - **A**, **T**, **G** или **C**, граничные эффекты устраняются замыканием последовательности нуклеотидов в кольцо.

Для словаря длины q максимальная возможная энтропия есть $\max(S_q) = q\ln(4)$.

Задача получения словаря длины $q+s$ из словаря длины q ставится и решается следующим образом [11]:

$$S_{q+s}\left[f_{q+s}\right] \rightarrow \max \quad (2)$$

при условии: если из восстановленного словаря длины $q+s$ получить суммированием словарь длины q, то он будет совпадать с исходным.

Решение (при $q>1$) получается методом множителей Лагранжа и может быть записано в явном аналитическом виде:

$$f_{i_1 \ldots i_q i_{q+1} \ldots i_{q+s}} = \frac{f_{i_1 \ldots i_q} f_{i_2 \ldots i_{q+1}} \cdots f_{i_{q-s+1} \ldots i_{q+s}}}{f_{i_2 \ldots i_q} f_{i_3 \ldots i_{q+1}} \cdots f_{i_{q-s+1} \ldots i_{q+s-1}}} \quad (3)$$

В знаменателе здесь стоят частоты словаря длины (q-1). Для q=1 метод максимума энтропии дает очевидное выражение $f_{i_1 \ldots i_{q+s}} = f_{i_1} \cdots f_{i_{q+s}}$.

Формула (3) представляет собой обобщение хорошо известных суперпозиционных приближений Кирквуда, Фишера и т.п. [6, 9, 10] из статистической физики, где они применяются для приближенного восстановления многочастичных функций распределения. Рассматриваемый случай примечателен тем, что полученные формулы восстановления являются точным решением поставленной задачи.

Следует особо подчеркнуть, что восстановленный словарь длины $q+s$ относится уже не к какому-либо однозначно определенному тексту, а к *ансамблю текстов*. У всех текстов этого ансамбля словари длины q совпадают с исходным словарем длины q.

4. Информативность и качество восстановления словарей

Восстановленный словарь, очевидно, должен быть более неопределенным, чем истинный словарь той же длины. Обозначим $S_{q+s}(q)$ энтропию словаря длины $q+s$, восстановленного по словарю длины q. Из формул (1) и (3) получаем очень простое выражение $S_{q+s}(q)$ через две энтропии S_q и S_{q-1} :

$$S_{q+s}(q) = (s+1)S_q - sS_{q-1} \quad (4)$$

где $q>1$, и $S_{1+s}(1) = (s+1)S_1$ для q=1.

Для обсуждения удобно использовать не энтропию, а *информативность* - величину $I_q = \max(S_q) - S_q = q\ln(4) - S_q$. Эта характеристика служит мерой упорядоченности. Нулевая информативность соответствует полной неопределенности (словарь содержит все слова, и все частоты равны между собой). Аналогично $S_{q+s}(q)$ определим $I_{q+s}(q)$ - информативность словаря длины $q+s$, восстановленного по словарю длины q.

Коэффициент восстановления $Q_{q+s}(q)$ есть отношение информативности восстановленного словаря к информативности истинного: $Q_{q+s}(q) = I_{q+s}(q)/I_{q+s}$. Оно лежит в пределах $0 \le Q_{q+s}(q) \le 1$. Равенство коэффициента восстановления единице соответствует случаю точного восстановления, а отличие от единицы характеризует различие истинного и восстановленного словарей.

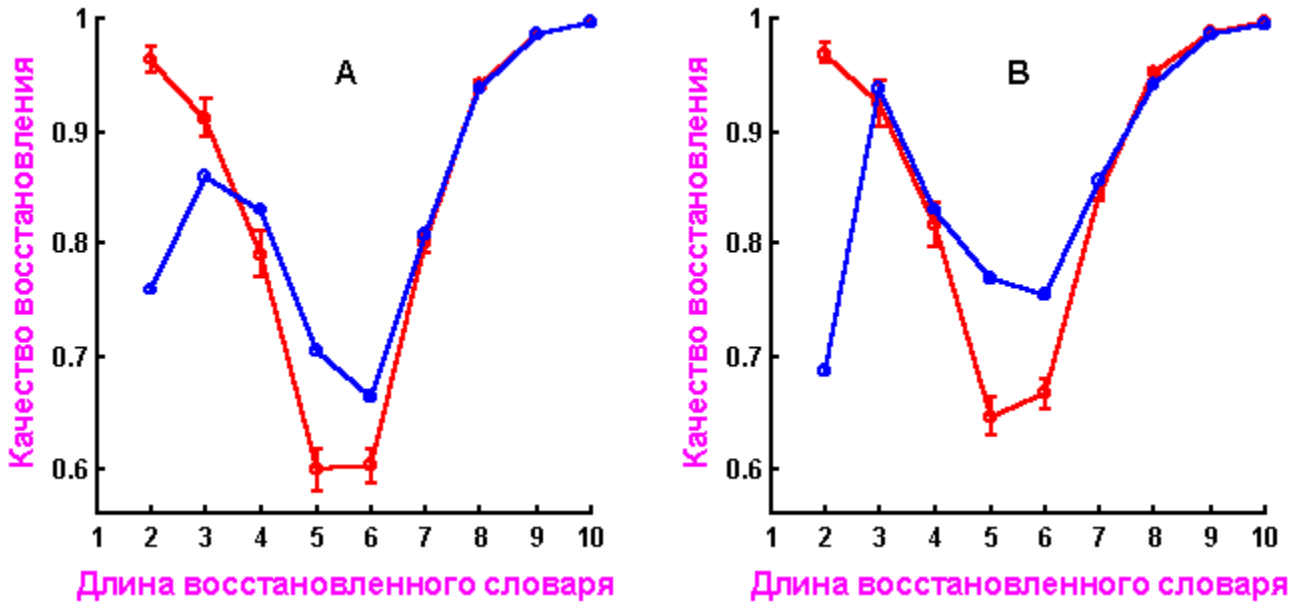
5. Сравнение реальных текстов со случайными.

Случайным текстом, соответствующим реальному, назовем текст той же длины с теми же пропорциями нуклеотидного состава, полученный по методу случайного выбора элементов.

Сравнивались реальные и соответствующие им случайные тексты. Для каждого текста были определены коэффициенты восстановления словаря данной длины по словарю на единицу меньшей длины. В машинном эксперименте определялись средние значения и средние квадратичные отклонения коэффициентов восстановления для случайных текстов.

Проанализированы банки генетических данных из тысяч различных нуклеотидных последовательностей. На рис.1 представлены типичные графики, получаемые в результате вычислений и последующих экспериментов со случайными последовательностями. Кривая средних для случайных текстов строилась по тридцати реализациям.

Проанализированы банки генетических данных из тысяч различных нуклеотидных последовательностей. На рис.1 представлены типичные графики, получаемые в результате вычислений и последующих экспериментов со случайными последовательностями. Кривая средних для случайных текстов строилась по тридцати реализациям.



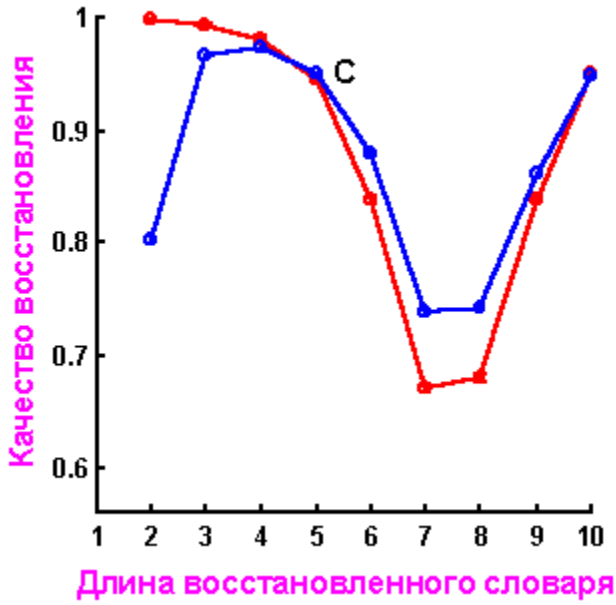


Рис. 1. Примеры зависимости качества восстановления частотных словарей по словарю на единицу меньшей длины от длины словаря.

A - последовательность из генома курицы (*N*=2136);

B - человека

(*N*=1639); C - нематоды (*N*=26139). По горизонтали откладывается длина восстановленных словарей, по вертикали - коэффициент восстановления. Для наглядности точки графиков соединены (для случайных текстов - сплошной красной линией, для реальных - сплошной синей линией), для случайных последовательностей показаны средние квадратичные отклонения

Характерно, что для реальных последовательностей словарь длины два восстанавливается по единичному словарю заметно хуже, чем для случайных текстов.

Качество восстановления словаря длины три по словарю длины два и длины четыре по длине три для реальных текстов уже достаточно велико и примерно такое же, как для случайных последовательностей, а словари длины 5 и 6 у реальных последовательностей восстанавливаются по предыдущим заметно (достоверно) лучше, чем у случайных. Дальнейшее поведение графиков коэффициентов восстановления, одинаковое для реальных и случайных текстов, обусловлено конечностью текстов. Регулярные достоверные отличия реальных текстов от случайных наблюдаются в машинных экспериментах для текстов достаточно длинных - как правило, более пятисот нуклеотидов.

6. Предельная энтропия как характеристика информационного содержания

Можно сравнивать словари одной длины различных текстов, сопоставляя их энтропии. Чтобы сравнивать словари разных длин, введем понятие *удельной предельной энтропии*..По данному словарю длины q восстановим словарь длины n>q и рассмотрим предел

$$S_\infty(q) = \lim_{n\to\infty}\left(S_n(q)/n\right) \quad (5)$$

Из формулы (4) получаем: $S_\infty(q) = S_q - S_{q-1}$ для q>1, и $S_\infty(1) = S_1$.

По смыслу своего определения это есть неопределенность, приходящаяся на одну букву бесконечно длинного текста, восстановленного по данному словарю длины q.

Удобно измерять удельную предельную энтропию в долях максимальной энтропии единичного словаря. При этом границы изменения устанавливаются от нуля (полная определенность) до единицы (полная неопределенность). Поэтому введем величину $L_\infty(q) = S_\infty(q)/\max(S_1)$,

которую далее будем просто называть *предельной энтропией*. Из (5) получаем $L_\infty(q) = \left(S_q - S_{q-1}\right)/\ln(4)$ для $q>1$, и $L_\infty(1) = S_1/\ln(4)$.

На рис.2 и рис.3 приведены графики предельных энтропий для реальных нуклеотидных последовательностей и соответствующих им случайных последовательностей.

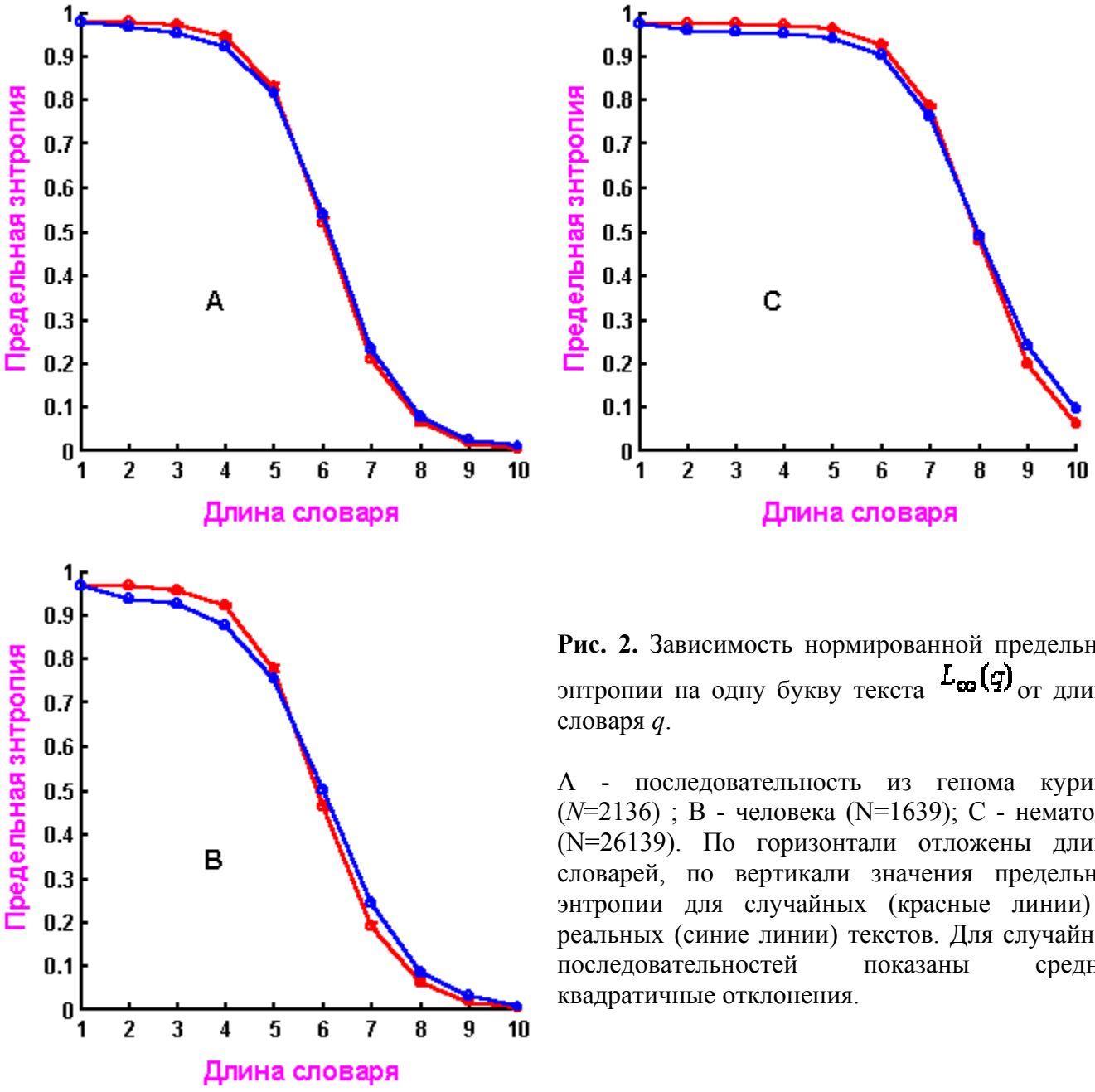


Рис. 2. Зависимость нормированной предельной энтропии на одну букву текста $L_\infty(q)$ от длины словаря q.

A - последовательность из генома курицы (N=2136) ; B - человека (N=1639); C - нематоды (N=26139). По горизонтали отложены длины словарей, по вертикали значения предельной энтропии для случайных (красные линии) и реальных (синие линии) текстов. Для случайных последовательностей показаны средние квадратичные отклонения.

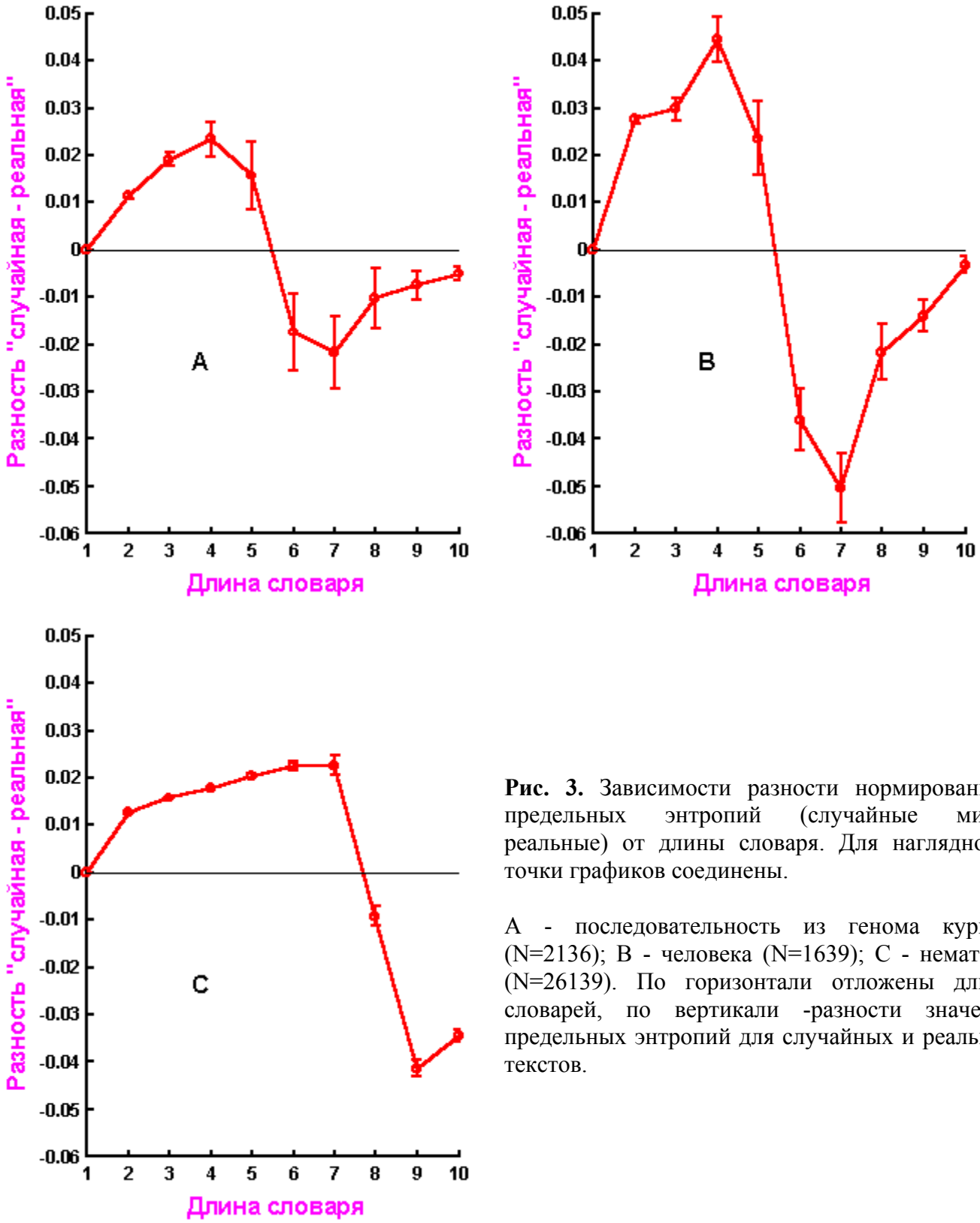


Рис. 3. Зависимости разности нормированных предельных энтропий (случайные минус реальные) от длины словаря. Для наглядности точки графиков соединены.

A - последовательность из генома курицы (N=2136); B - человека (N=1639); C - нематоды (N=26139). По горизонтали отложены длины словарей, по вертикали -разности значений предельных энтропий для случайных и реальных текстов.

Разность предельных энтропий $L_\infty(q)-L_\infty(q+1)$ - двух соседних словарей характеризует прирост информации о тексте при переходе от словаря длины q к словарю длины q+1. **Среди таких переходов при q=1:9 для исследованных реальных текстов наименьшее изменение предельной энтропии наблюдается при переходе от q=2 к q=3, а для случайных - при переходе от q=1 к q=2.** Это означает, что, хотя информация в реальных последовательностях кодируется тройками (кодонами), значительная ее часть содержится на длине два.

Для *q*=2:5 словари реальных генетических последовательностей достоверно содержат больше информации о тексте, чем словари случайных последовательностей.

В подавляющем большинстве случаев более 90% информации о гене содержится в его частотном словаре длины 8.

При помощи предельной энтропии мы получаем возможность сравнивать информационное содержание словарей различной длины одного текста и сопоставлять между собой словари разных текстов, независимо от длины как словарей, так и текстов.

7.Заключение

Итак, построен метод информационного анализа генетических текстов, основанный на сравнении энтропии и информационных характеристик частотных словарей. Проведен анализ реальных нуклеотидных последовательностей и выделены их достоверные отличия от случайных.

Представляется перспективным использование частотных словарей малой длины *q*=2:8 для сравнительного и кластерного анализа реальных генетических текстов.

Работа проводилась при частичной поддержке Красноярского краевого фонда науки, гранты 3F0190 и 4F0153.

Литература

1.Компьютерный анализ генетических текстов/ Александров А.А., Александров Н.Н., Бородовский М.Ю. и др. М.: Наука, 1990. 267 с.
2. Garden P.W. Markov Analysis of Viral DNA/RNA sequences// J.Theor. Biol. , 1980. Vol. 82. P. 679 - 684.
3.Brendel V., Beckmann J.S., Trifonov E.N. Linguistics of nucleotide sequences: morphology and comparison of vocabularies// J.Biomol.Struct.Dyn., 1986. Vol.4. P. 11 -22.
4. deWachter R. The Number of Repeats Expected in Random Nucleic Acids Sequences and Found in Genes// J. Theor. Biol., 1981. Vol. 91. P. 71-98.
5.Pevzner P.A., Borodovski M.Yu., Mironov A.A. 1. The significance of deviation from mean statistical characteristics and prediction of the frequency of occurences of words// J.Biomol. Struct. Dyn., 1989. Vol. 6. P. 1013 - 1026.
6. Исихара А. Статистическая физика - М.: "Мир", 1973, 466с.
7.Горбань А.Н., Миркес Е.М., Попова Т.Г., Садовский М.Г. Новый подход к изучению статистических свойств генетических последовательностей// Биофизика, 1993. Т.38. Вып.5. С.762 - 767.
8.Горбань А.Н., Попова Т.Г., Садовский М.Г. Избыточность генетических последовательностей и мозаичная структура генома// Мол. биология, 1994. Т.28. Вып.2. С. 313 - 324.
9. Бугаенко Н.Н., Горбань А.Н., Карлин И.В. Универсальное разложение трехчастичной функции распределения// Теорет. и мат. физика, 1991. Т.88, ©3. С.430-441.
10. Бугаенко Н.Н., Горбань А.Н., Карлин И.В. Универсальная зависимость $F_3[F_2]$. // Термодинамика необратимых процессов. М.: Наука, 1992. С.30-38.
11. Бугаенко Н.Н., Горбань А.Н., Садовский М.Г. Об определении информационной емкости нуклеотидных последовательностей// Мол. биология, 1996. Т.30, вып. 3. С. 252-268.