Исследование повторов в текстах. Формальные языковые системы

Гусев Владимир Дмитриевич, к.т.н., ИМ СО РАН

Кафедра информационной биологии ФЕН НГУ

Содержание лекции 1.

1. Естественные и формальные языковые системы.
    - 1.1. Трактовка языка в широком смысле.
    - 1.2. Естественный язык: примеры определений и основные признаки.
    - 1.3. Формальные языки.
2. Межъязыковые аналогии.
    - 2.1. Определение расстояния между цепочками символов.
    - 2.2. Поиск потенциально возможных разделителей в АМ-последовательностях.
    - 2.3. Локализация искажений в несовершенных повторах.
3. Классификация повторов.
    - 3.1. По способу их прочтения и использованию переименований.
    - 3.2. По наличию искажений.
    - 3.3. По характеру расположения в тексте.
4. Примеры повторов в генетических текстах.
5. Представление текста в терминах повторов.
    - 5.1. Полный частотный спектр текста.
    - 5.2. Меры сходства

Вопросы, упражнения, задачи к лекции 1.

1. Естественные и формальные языковые системы.

1.1. Трактовка языка в широком смысле.

Единого и достаточно содержательного определения понятия "язык", по-видимому, не существует. Одна из причин – быстрый рост числа объектов, к которым применяется этот термин ("язык пчел", "язык дельфинов", "язык программирования", "язык запросов" (в информационных системах), "генетический язык" и др.).

С точки зрения лингвиста:

"**Язык** – это любая знаковая система". Знак – это условное обозначение некоторого предмета, свойства, отношения, явления.

Знаковая система – совокупность знаков, организованная определенным образом.

Семиотика – наука, изучающая свойства знаков и знаковых систем (как естественных, так и формальных).

С точки зрения математика:

Пусть

Σ – непустое конечное множество символов (алфавит);

цепочка ***W*** (слово или строка) – конечная последовательность символов из Σ;

$|W|$ – длина цепочки W;

пустая цепочка e – это цепочка. не содержащая символов ($|e| = 0$);

Σ* – множество всех цепочек в алфавите Σ, включая e.

Язык *L* над алфавитом Σ – произвольное множество цепочек в Σ (т.е. $L \subseteq \Sigma^*$).

1.2. Естественный язык: примеры определений и основные признаки.

Более содержательными являются определения конкретных языков, например, естественного:

" **Язык** это система звуковых и словарно-грамматических средств, являющаяся орудием для выражения мысли и служащая средством общения людей" ("Словарь русского языка", т. IV, М., 1961).

" **Язык** – это стихийно возникающая в человеческом обществе и развивающаяся система дискретных (членораздельных) звуковых знаков, предназначенная для целей коммуникации и способная выразить всю совокупность знаний и представлений человека о мире" (БСЭ).

"**Язык** – одна из самобытных семиологических систем, являющаяся основным и важнейшим средством общения членов данного человеческого коллектива, для которых эта система оказывается также средством развития мышления, передачи от поколения к поколению культурно-исторических событий и т.п. " (О.С. Ахманова. Словарь лингвистических терминов).

Основные признаки, характеризующие представления экспертов о языке:

1) язык – средство коммуникации и выражения мыслей (функциональный аспект);

2) стихийность возникновения и развития (эволюционный аспект);

3) дискретность (внутренняя членимость сообщения);

4) иерархичность (комбинирование более крупных единиц из исходных мелких: морфемы – слова – предложения и т.д.);

5) открытость (возможность составления большого (потенциально бесконечного) числа осмысленных сообщений).

дополнительные (реже упоминаемые) признаки:

6) отсутствие жесткой однозначной связи между обозначением и обозначаемым (один и тот же объект может иметь несколько обозначений (явление синонимии) и одно и то же обозначение может быть использовано для разных объектов (явление омонимии));

7) экономичность кода (длина сообщения должна быть пропорциональна количеству информации в этом сообщении).

1.3. Формальные языки

чаще всего задаются в виде некоторой схемы порождения (порождающей или формальной грамматики), которая позволяет получить все цепочки данного языка и только их. Обычно порождающая грамматика представляется в виде четверки

$G = (\Sigma, N, P, S)$, где

Σ – алфавит терминальных символов, из которых составляются "предложения" языка ($L(G) \subseteq \Sigma^*$);

N – алфавит нетерминальных символов (или переменных);

$\Sigma \cap N = \varnothing$;

P – конечное множество правили вывода вида $\alpha \to \beta$, где $\alpha \in (N \cup \Sigma)^* N (N \cup \Sigma)^*$, $\beta \in (N \cup \Sigma)^*$;

S – выделенный символ из N, называемый начальным (или исходным).

Формальные грамматики были введены Хомским (1956г). Им же была построена иерархия грамматик 4 типов.

Пример 1.

Пусть $G = (\{a,b,c\}, \{A,B,S\}, P, S)$,

где правила вывода P имеют вид:

$S \to AB$, $A \to a$, $A \to ac$, $B \to b$, $B \to cb$.

Данная грамматика позволяет получить всего 4 вывода терминальных строк:

(1) $S \to AB \to aB \to ab$

(2) $S \to AB \to aB \to acb$

(3) $S \to AB \to acB \to acb$

(4) $S \to AB \to acB \to accb$

$L(G) = \{ab, acb, accb\}$.

Для строки acb имеются два разных вывода.

Пример 2. (язык образцов, формально не вкладывающийся в иерархию Хомского).

Пусть

$\Sigma = \{a,b,c,d\}$ – алфавит константных символов,

$N = \{X, Y\}$ – алфавит переменных.

Образец **R** = X^2aY^2bXYcd (цепочка из константных и переменных символов).

Правила вывода: вместо X и Y могут быть подставлены любые непустые цепочки из Σ^* (вхождениям одной и той же переменной соответствует одна и та же цепочка).

Языку **L(R)** принадлежат цепочки $d^2ac^4bdc^3d$

(соответствующие подстановки имеют вид: $X \to d$; $Y \to cc$),

$a^7b^3a^3bcd$ (подстановка: $X \to a^3$; $Y \to b$) и др.

Пример 3 (язык образцов).

Пусть $\Sigma = \{A, G, C, T\}$ – алфавит ДНК-последовательностей; $N = \{X, Y\}$;
правила вывода:

$X \to \alpha, \quad \alpha \in \Sigma^*, \quad |\alpha| \geq k_1; \qquad Y \to \beta, \quad \beta \in \Sigma^*, \quad |\beta| \leq k_2,$

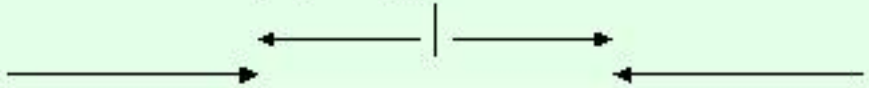
$R = XYY^{сим}\, X^{сим,\, комп.},$

где $Y^{сим}$ – цепочка симметричная по отношению к Y;

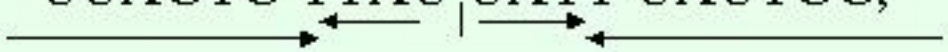
$X^{сим,\, комп.}$ – цепочка, получающаяся из X изменением порядка следования элементов на противоположный и комплементарной подстановкой ($A \leftrightarrow T, \; C \leftrightarrow G$).

Пусть для определенности $k_1 = 6$, $k_2 = 4$.

Тогда образец R задает структуры шпилечного типа с симметричной петлей:



CCAGTG TTAC CATT CACTGG,



CT CT CT C TAT TAT G AG AG AG и т.п.

Основной вывод:

генетический язык по своим свойствам
(стихийность возникновения,
способность эволюционировать,
иерархичность,
наличие синонимии и омонимии)
ближе к естественным языкам, чем к формальным.
Под вопросом лишь «экономичность» кода.

2. Межъязыковые аналогии.

2.1. Определение расстояния между цепочками символов.

За период с 1965 по 1975гг, по-видимому независимо, и под разными наименованиями для разных языковых систем были введены близкие по смыслу понятия расстояния. В простейшем виде расстояние d(u,v) между цепочками u и v определяется как минимальное число операций типа "вставка", "устранение" или "замена" символа, переводящих одну цепочку в другую.

Пример: u = abbaeac; v = bdedac

u → v:

a	b	b	a	e	a	c
	|				|	|
–	b	d	e	d	a	c

d(u,v) = 4 (1 дел., 3 зам).

u → v:

a	b	b	a	e	–	a	c
	|			|		|	|
–	b	d	–	e	d	a	c

d(u,v) = 4 (2 дел., 1 вст., 1 зам).

Хроника введения расстояний и мер сходства, основанных на идеях динамического программирования.

Год	Наименование (меры, расстояния, процедуры сравнения)	Авторы	Предметная область
1965	Метрика Левенштейна	Левенштейн В.И.	теория связи (двоичные коды)
1968	процедуры нелинейной нормализации речи по темпу	Слуцкер Г.С.	распознавание речи
1968	ДП - метод	Винцюк Т.К.	распознавание речи
1969	дискретный вариант меры Слуцкера	Загоруйко Н.Г., Величко В.М.	распознавание речи
1970	мера сходства АМ-последовательностей	Needleman S.B., Wunch S.B.	молекулярная биология
1974	редакционное расстояние	Wagner R.A., Fisher M.J.	лингвистика
1974	эволюционное расстояние	Sellers P.	молекулярная биология

2.2. Поиск потенциально возможных разделителей в АМ-последовательностях.

Критерий:

а) отсутствие тандемных вхождений;

б) "сверхравномерность" позиционного распределения (максимальное расстояние между соседними вхождениями аминокислоты в текст слишком мало, а минимальное – слишком велико) – запрет на излишнее сближение и удаление.

Пример обработки нейроминидазы (NA) вируса гриппа:

"сверхравномерное" распределение демонстрирует глицин (F = 44).

Пример обработки текста "Винни-Пуха":

"сверхравномерное" распределение демонстрирует слово "глава".

Гипотеза Трифонова о "сверхравномерном" распределении метионина (Met) в АМ-последовательностях.

2.3. Локализация искажений в несовершенных повторах.

Пусть u и v – два слова естественного языка близких в смысле редакционного расстояния.

Для определенности рассмотрим случай, когда

$d(u,v) = 2$ и $d(u,v) << \min(|u|, |v|)$

(достаточно длинные слова с двумя отличиями:

SS – замена + замена, IS – вставка + замена, I I -- две вставки и т.п.) .

Экспериментально (на 100-тысячном словаре) зафиксирован следующий факт: число позиционно кластеризованных искажений в близких словах существенно выше уровня, допускаемого моделью с независимым распределением искажений по длине слова.

Объяснение: совместно искажаемые позиции принадлежат структурной единице более низкого уровня (слог, морфема).

Примеры:

а) $d(u,v) = 2$ – замены слогов (SS), вставки морфем (I I):

комб.-я операций	*u*		*v*
SS:	синеватый	–	сизоватый, сиповатый, (но!) виноватый;
I I :	скрыть	–	скрывать, скрыться, раскрыть;

б) $d(u,v) = 1$ – в длинных словах единичные замены, вставки, делеции часто возникают на границах структурных единиц.

тип операции	*u*		*v*
I	франко-русский	–	франко-прусский
	двустворчатый	–	двухстворчатый
S	склониться	–	склоняться
S	перепариваться	–	пережариваться, перевариваться

Гипотеза (по аналогии): анализ локализации искажений (как единичных, так и множественных кластеризованных) в достаточно длинных несовершенных повторах, выявляемых в анализируемом тексте, может быть полезен для установления структурных единиц этого текста.

3. Классификация повторов.

3.1. По способу их прочтения и использованию переименований:

Повторы – элементарные структурообразующие компоненты текстов.

Будем называть повтором (в широком смысле) пару фрагментов текста, совпадающих с точностью до переименования элементов алфавита и (возможно), изменения направления считывания элементов в одном из фрагментов на противоположное.

В соответствии с этим будем различать следующие типы повторов:

– прямые : … AGTTC … AGTTC…

– симметричные: … AGTTC … CTTGA…

– с точностью до подстановки на элементах алфавита: секвентные переносы в музыкальных произведениях; замены $0 \to 1$ и $1 \to 0$ в двоичных словах; замены $A \to T$, $T \to A$, $C \to G$, $G \to C$, связанные с отношением комплементарности в ДНК-последовательностях:

– прямые комплементарные: … AGTTC… TCAAG…

– симметричные комплементарные: … AGTTC… GAACT…

3.2. По наличию искажений:

Повторы могут быть

совершенные:

… AGTTC … AGTTC…

и несовершенные (с заменами, вставками, делециями):

… AGTTC … AATTC … (замена),

… AGTTC … AGTTTC… (вставка),

в том числе с точностью до агрегирования:

… AGTTC … GATCT …

(совпадают при заменах {A,G} → Pu, {C,T} → Py).

3.3. По характеру расположения в тексте:

будем различать повторы

разнесенные

… AGTTC … AGTTC…

тандемные … AGTTC AGTTC…

с наложением : … AGTTCAGTTCAGTTC …

4. Примеры повторов в генетических текстах

1) Участки с аномальным нуклеотидным составом:

   (A,T)-богатые, (C,G)-богатые и т.п.

   Тип повторов – как правило, несовершенные.

2) Участки микросателлитной ДНК: периодичности с малой длиной периода (обычно 2,3) и достаточно высокой кратностью повторений ($(TA)^n$, $(CAG)^n$… n – порядка 10 и выше). Тип повторов: обычно совершенные.

3) Тандемные повторы со средней и большой длиной периода (L ~10 и выше) и кратностью повторений, уменьшающейся "в среднем" с увеличением длины периода. Тип повторов: совершенные и несовершенные.

4) Разнесенные повторы значительной длины (L ~10^2), совершенные (например, длинные концевые повторы в геномах некоторых микроорганизмов) и несовершенные (повторы крупных блоков в кодирующих участках некоторых генов).

5) Повторы регуляторных и функциональных единиц (как правило, несовершенные):

– рибосомные сайты связывания, промоторы и терминаторы транскрипции у прокариотов (L ~$10 - 10^2$);

– сайты связывания транскрипционных факторов у эукариотов (L ~$10 - 20$);

– гены и псевдогены (L ~$10^3 - 10^4$);

– повторы отдельных генов (рибосомальные, гистоновые…), L ~$10^3 - 10^4$;

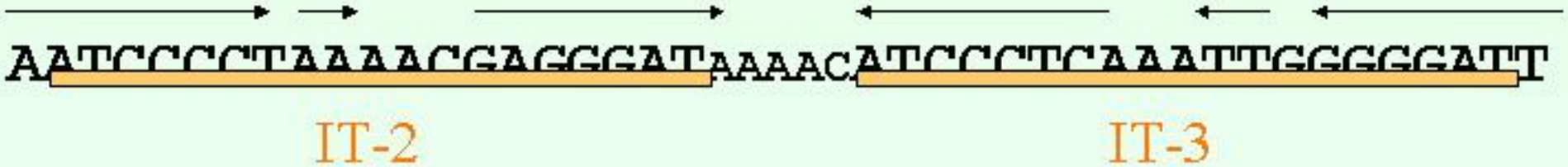
– мобильные элементы (L ~$10^3 - 10^4$);

6) Блоки из регуляторных и функциональных единиц с вкраплением некодирующих участков, часто соответствующих отдельным транскриптам ($L \sim 10^4$);

Следует отметить, что многие регуляторные и функциональные единицы, в свою очередь, построены по принципу тандемной повторности. Приведем для иллюстрации выравнивание, соответствующее тандемным повторам ("итеронам"), расположенным в зоне начала репликации бактериофага λ:

IT-1:	39034:	ATCCCTCAAAACGAGGGAA	A ↵
IT-2:	39054:	ATCCCCTAAAACGAGGGAT	AAAAC ↵
IT-3:	39078:	ATCCCTCAAATTGGGGGAT	TGCT ↵
IT-4:	39101:	ATCCCTCAAAACAGGGGGA	

Здесь значок ↵ означает, что нижеследующая строка непосредственно продолжает предыдущую. Интересно отметить своего рода эффект "локальной фрактальности": палиндромно-шпилечные структуры внутри итеронов приводят к возникновению аналогичных (но более сильных) структур между итеронами:



5. Представление текста в терминах повторов.

5.1. Полный частотный спектр текста. Пусть

Σ – конечный алфавит;

S – текст, составленный из элементов Σ;

$N = |S|$ – длина текста;

$S[i]$ – i-й элемент текста S $(1 \le i \le N)$;

$S[i:j]$ – фрагмент текста, включающий элементы с i-го по j-й $(1 \le i < j \le N)$.

Назовем l-граммой связную цепочку текста, содержащую l символов.

Всякий текст $S = a_1 a_2 a_3 a_4 a_5 \ldots a_N$

можно представить в виде последовательности l-грамм ($l = 1,2,\ldots,N$), выделяемых скользящим окном ширины l, движущимся вдоль текста с шагом в 1 символ.

Полное число l-грамм равно $N - l + 1$.

С учетом возможных повторений число различных l-грамм $M_l \le N - l + 1$.

Назовем *частотной характеристикой l-го порядка* текста S совокупность элементов $\Phi_l(S) = \{\phi_{l1}, \phi_{l2}, \ldots, \phi_{l M_l}\}$

где $\phi_{l M_i}$ $(1 \le i \le M_l)$ есть пара :

<*i*-я *l*-грамма – x_i, частота ее встречаемости в тексте – $F_l(x_i)$ >.

Пусть $l_{max}(S)$ – наибольшее значение l, при котором в тексте S еще содержатся повторяющиеся l-граммы.

Совокупность частотных характеристик

$$\Phi(S) = \{\Phi_1(S), \Phi_2(S), \ldots, \Phi_{l max + 2}(S)\}$$

назовем *полным частотным спектром текста S*.

По нему текст S может быть восстановлен однозначно.

На практике чаще используют усеченный спектр

$$\Phi^*(S) = \{\Phi^*_1(S), \Phi^*_2(S), \ldots, \Phi^*_{l max}(S)\}$$

куда не входят $\Phi_{l max+1}(S)$ и $\Phi_{l max+2}(S)$, а $\Phi^*_l(S)$ отличаются от $\Phi_l(S)$ лишь отсутствием l-грамм с единичной частотой встречаемости.

Пример. Пусть S = caabcabbca.

Тогда $\Phi^*(S) = \{\Phi^*_1(S), \Phi^*_2(S), \Phi^*_3(S)\}$, где

$\Phi^*_1(S) = \{$<a, F (a) = 4>; <b, F(b)= 3>; <c, F(c) = 3>$\}$;

$\Phi^*_2(S) = \{$<ca, F(ca) = 3>; <ab, F(ab)= 2>; <bc, F(bc) = 2>$\}$;

$\Phi^*_3(S) = \{$<bca, F(bca) = 2>$\}$;

Позиционную информацию целесообразно выдавать лишь для повторов значительной длины.

Наиболее важными являются следующие параметры частотного спектра:

$l_{\max}$ — длина максимального повтора в тексте.

Для случайного текста длины N с вероятностями появления элементов алфавита p_r $(1 \le r \le n = |\Sigma|)$ можно пользоваться следующей оценкой: .

$$l_{\max} \sim 2\ln N \Big/ \left|\ln \sum_{r=1}^{n} p_r^2\right|$$

Если реальная длина $l_{\max}$ в тексте существенно превышает ожидаемое значение, это свидетельствует о наличии дупликативных механизмов порождения текста.

~~M_l~~ — размер словаря l-грамм.

Он фигурирует в определении комбинаторной сложности текста.

$F_l^{\max} = \max\limits_{1 \le i \le M_l} F_l(x_i)$ — максимальное значение частот встречаемости l-грамм в тексте $(1 \le l \le l_{\max})$;

$F_l^{\min} = \min\limits_{1 \le i \le M_l} F_l(x_i)$ — минимальное значение частот встречаемости

l –грамм в тексте $(1 \le l \le l_{\max})$; представляет интерес

лишь при малых значениях l; при больших обычно $F^{\min} = 1$

E_l^k — Число различных l-грамм, каждая из которых встречается в тексте ровно k раз ($k = 1,2,\ldots,N-l+1$); зависимость E_l^k от k при фиксированном l является аналогом известной в лингвистике кривой Юла;

$M_l = E_l^1$ — число различных повторяющихся l-грамм;

E_l^0 — число l-грамм, ни разу не встретившихся в тексте. Наличие таковых в ситуации, когда $N / n^l >> 1$, можно трактовать как аномальный эффект.

Имеют место простые соотношения, связывающие основные параметры:

$$M_l = \sum_{k=1}^{F_l^{\max}} E_l^k, \qquad N - l + 1 = \sum_{k=1}^{F_l^{\max}} k \cdot E_l^k, \qquad E_l^0 = n^l - M_l.$$

С помощью частотного спектра текста можно:

а) сравнить реальный текст с его "случайной копией" с тем же составом элементов ($\Phi_1(T)$), но с разрушенными связями между символами; выявить при этом аномально длинные повторы, аномально частые и аномально редкие цепочки, цепочки, сохраняющие частоту при расширении

(см. переход от ***q*** к ***qu*** в английском) и т.п.;

б) получить оценки переходных вероятностей при построении марковских моделей текста

(необходимо, чтобы размер текста был достаточно велик);

в) вычислить энтропийные характеристики l-го порядка и оценки избыточности текста (Шеннон);

г) получить численные оценки близости двух (или большего числа) текстов в ситуации, когда перестают работать методы выравнивания.

5.2. Меры сходства (иллюстрация п. "г").

а) Пусть T_1 и T_2 два текста.

Назовем совместной частотной характеристикой l-го порядка текстов T_1 и T_2 совокупность элементов

$$\Phi_l(T_1, T_2) = \{\phi_{l1}(T_1, T_2), \phi_{l2}(T_1, T_2), \ldots, \phi_{lM_l}(T_1, T_2)\}$$

где $M_l = M_l(T_1, T_2)$ — число l-грамм, общих для обоих текстов,

а элемент ϕ_{li} $(1 \le i \le M_l)$ есть тройка:

<< i-я общая l-грамма – x_i,

частота ее встречаемости в T_1 – $F(T_1, x_i)$ и в T_2 – $F(T_2, x_i)$ >>.

Простейший набор мер сходства, упорядоченный по возрастанию l имеет вид:

$$q_l(T_1, T_2) = \frac{M_l(T_1, T_2)}{M_l(T_1) + M_l(T_2) - M_l(T_1, T_2)},$$

$l = 1, 2, \ldots, l_{\max}(T_1, T_2)$

б) более сложный вариант, учитывающий частоты встречаемости *l*-грамм:

$$\lambda(T_1,T_2)=\frac{\sum_{\alpha}\min\{F(T_1,\alpha),F(T_2,\alpha)\}\cdot|\alpha|}{\sum_{\alpha}\max\{F(T_1,\alpha),F(T_2,\alpha)\}\cdot|\alpha|}$$

где α – произвольная цепочка текстов T_1 и (или) T_2,

$|\alpha|$ – ее длина.

(Findler N.V., Van Leeuwen, 1979)

в) пример расстояния (Kohonen T., Reyhkala E., 1978)

$$h(T_1, T_2) = \max(m(T_1), m(T_2)) - m_0(T_1, T_2) \qquad (*)$$

где $m(T_1)$и $m(T_2)$ – количество признаков, выделяемых из текстов T_1 и T_2, а $m_0(T_1, T_2)$ — число совпавших признаков. Если в качестве признаков использовать l-граммы, получаемые сдвигом вдоль текста скользящего окна размера l, то $m(T_1) = |T_1| - l + 1$, $m(T_2) = |T_2| - l + 1$, и l-граммный аналог (*) имеет вид: .

$$h_l(T_1, T_2) = \max(|T_1| - l + 1, |T_2| - l + 1) - \sum_{i=1}^{M_l(T_1, T_2)} \min\{F(T_1, x_i), F(T_2, x_i)\}$$

г) ранговая мера близости:

Пусть l-граммы в $\Phi_l(T_1)$ и $\Phi_l(T_2)$ упорядочены по убыванию частот; порядковое место l-граммы x_i в упорядочении определяет ее ранг – $r(T_1, x_i)$ (соответственно, $r(T_2, x_i)$).

Группы равночастотных l-грамм представляются усредненным рангом.

Введем l-граммный аналог расстояния:

$$S_l(T_1,T_2) = \sum_{x_i \in \Sigma_l} \left(r(T_1,x_i) - r(T_2,x_i)\right)^2$$

где Σ_l – совокупность всевозможных цепочек длины l ; $|\Sigma_l| = R_l = n^l$.

Аналогом коэффициента Спирмэна для характеристики l-го порядка ($l = 1,2,\ldots$) является

$$\rho_l(T_1,T_2) = 1 - \frac{6S_l(T_1,T_2)}{R_l(R_l^2 - 1)} \qquad (**)$$

При наличии равночастотных l-грамм в (**) вносится поправка на "связанность" рангов. (Кендел М. Ранговые корреляции, М., Статистика, 1975)

Вопросы, упражнения, задачи.

1. Привести примеры синонимии и омонимии в генетических текстах. Существуют ли аналоги антонимов?
2. а) Синтезируйте цепочку нуклеотидов минимальной длины, которая перекрывала бы терминальными кодонами все 3 допустимых рамки считывания.

   б) то же самое с дополнительным ограничением: цепочка должна представлять собой тандемный повтор.

   Единственное ли решение?
3. Синтезируйте цепочку нуклеотидов минимальной длины, которая кодирует цистеин (хотя бы одно вхождение) в любой из трех рамок считывания.
4. Приведите пример аминокислоты, кодируемой триплетами, некоторые из которых образуют прямые комплементарные повторы.
5. В порождающей грамматике $G = (\Sigma, N, P, S)$ алфавит терминальных символов $\Sigma = \{a,b,c\}$, нетерминальных – $N = \{S,T\}$, S – начальное состояние, а правила вывода P имеют вид: $S \to aT$, $T \to bT$, $T \to c$.

   Описать общий вид цепочек, порождаемых G.

6. Определите, являются ли порождающими следующие грамматики:

а) $G = (\Sigma = \{a,b\};\quad N = \{S\};\quad P = \{S \to aSbaS,\quad S \to ab,\quad aS \to B\};\quad S)$;

б) $G = (\Sigma = \{b,c\};\quad N = \{A, S\};\quad P = \{S \to A,\quad A \to b,\quad A \to cA\};\quad S)$;

7. Первые два куплета русской народной песни "Я на камушке сижу" имеют вид:

1-й куплет	2-й куплет
Я на камушке сижу,	Я топор в руке держу,
Я топор в руке держу,	Вот я колышки тешу,
Ай ли, ай люли,	Ай ли, ай люли,
Я топор в руке держу.	Вот я колышки тешу.

Представьте структуру куплетов этой песни в виде образца (шаблона), содержащего константные (не меняющиеся от куплета к куплету) компоненты и переменные, допускающие различные подстановки в разных куплетах.

8. Представьте на языке образцов структуру куплетов следующей песенки:

One man went to mow, went to mow a meadow,
One man and his dog, went to mow a meadow,
Two men went to mow, went to mow a meadow,
Two men,
One man and his dog, went to mow a meadow,
Three men went to mow, went to mow a meadow,
Three men, Two men,
One man and his dog, went to mow a meadow,
…
Nine men went to mow, went to mow a meadow,
Nine men, Eight men, Seven men, Six men,
Five men, Four men, Three men, Two men,
One man and his dog, went to mow a meadow!

9. Каким единым образцом описываются следующие фрагменты АМ-последовательностей «Adenomatous polyposis coli protein» ($N = 2845$ а.о.):

   1) PGRNSISPGRN;
   2) SESDSSE;
   3) SNGNGSN;
   4) SEKAKSE

10. Сформулируйте условие, при котором комплементарный палиндром одновременно реализует прямой комплементарный повтор.

11*. Покажите, что редакционное расстояние удовлетворяет метрическим аксиомам.

12. Укажите, не менее 5 вариантов внесения двукратных искажений типа "делеция символа" (D), и "вставка" (I) в слово "сорванный", приводящих к осмысленному слову. Есть ли варианты изменения редакционных операций, приводящие к тому же решению.

13. Постройте ориентировочный график «устойчивости» различных позиций в слове «заживать» к одноэлементным «осмысленным» заменам. Как объяснить результат?

14. Проиллюстрируйте примеры структур, выявляемых в следующих последовательностях:

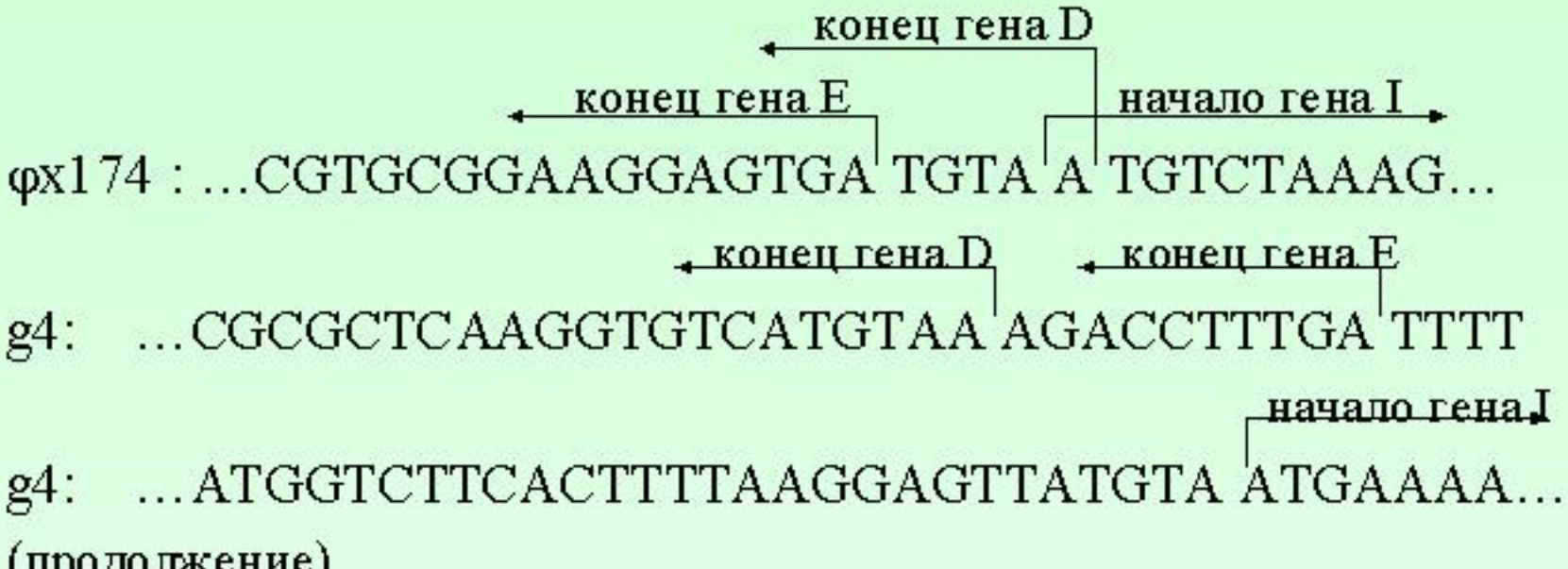
а) AACAGCACCACGGTGGTGGTGAACACGGTGG

– фрагмент orf-401 из генома бактериофага λ.

б) CTGGCCACAACCCCCCACTGGCCAGGCCGTCCCTCCCCACTGGCCCT

(фрагмент эукариотического промотора HSLCATG, начало – в (-92)-й позиции от точки инициации транскрипции).

15*. На нижеследующем рисунке представлены близкие в функциональном отношении фрагменты из геномов родственных бактериофагов (φx174 и g4)



(продолжение)

Опишите характер эволюционных перестроек, имевших место в данном районе геномов.