ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Октябрь—декабрь 1999. Серия 1. Том 6, № 4, 92–103

УДК 519.114

О ГРАФАХ ПОДСЛОВ DOL-ПОСЛЕДОВАТЕЛЬНОСТЕЙ*)

А. Э. Фрид

Получено описание графов подслов, или графов Рози, для равноблочных маркированных циркулярных DOL-последовательностей.

Введение

С каждой последовательностью над конечным алфавитом можно связать семейство $\{G(n)\}$ *графов Рози́*, или *графов подслов* длины n. Введенные в [10] графы Рози активно используются для изучения символьных последовательностей.

В настоящей работе получено описание графов Рози для широкого семейства последовательностей, в котором содержатся многие известные примеры, включая последовательность Туэ–Морса. Показано, что чтобы узнать вид графа Рози последовательности этого класса для сколь угодно большой длины, достаточно конечным перебором найти графы Рози для длин, ограниченных некоторой константой. Тем самым продолжено исследование, начатое в работах [1, 5] и [6], в которых были получены аналогичные результаты для двух других свойств последовательностей из данного класса — комбинаторной сложности и частоты слов.

1. Предварительные определения

Конечное слово u называется *подсловом* конечного или бесконечного слова v над алфавитом Σ, если $v = s_1us_2$ для некоторых слов s_1 и s_2.

Пусть w — последовательность над алфавитом Σ. Обозначим через F множество всех ее подслов; множество всех подслов из F длины n обозначим через $F(n)$. Ниже рассматриваем случай, когда язык F *продолжаем*, т. е. для любого слова $v \in F$ найдутся такие символы $a, b \in \Sigma$, что $avb \in F$.

*) Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (коды проектов 99–01–00531 и 97–01–01075) и Федеральной целевой программы «Интеграция» (проект 1997–473).

Правой степенью специальности $r(u)$ слова $u \in F$ называется число букв из Σ, которыми u может быть продолжено вправо до подслова последовательности w. Аналогично вводится понятие *левой степени специальности* $l(u)$.

Типом слова $u \in F$ назовем пару $(l(u), r(u))$.

Слово u называется *специальным справа*, если $r(u) \neq 1$, и *специальным слева*, если $l(u) \neq 1$. Слово будем называть *ординарным*, если оно не является специальным ни слева, ни справа. Наконец, слово называется *биспециальным*, если оно является одновременно специальным и слева, и справа.

Комбинаторной сложностью $f(n)$ последовательности w называется число подслов в w длины n: $f(n) = |F(n)|$. Вводится также функция $s(n)$ первых разностей комбинаторной сложности: $s(n) = f(n+1) - f(n)$.

Типом вершины ориентированного графа назовем пару, состоящую из его полустепени захода и полустепени исхода. Вершину будем называть *ординарной*, если ее тип равен $(1, 1)$, и *специальной* в противном случае.

Маршрут ненулевой длины в орграфе G назовем *прямым*, если первая и последняя вершины этого маршрута специальны, а все остальные вершины ординарны.

Схемой графа G назовем граф, в котором вершинами являются специальные вершины графа G, а ребра соответствуют всем прямым маршрутам в графе G. Два графа называются *гомеоморфными*, если изоморфны их схемы. Гомеоморфность графов G_1 и G_2 будем обозначать через $G_1 \simeq G_2$.

2. DOL-последовательности

Морфизмом над алфавитом $\Sigma = \{a_1, \dots, a_q\}$ называется отображение $\varphi : \Sigma^* \to \Sigma^*$, сохраняющее конкатенацию, т. е. $\varphi(xy) = \varphi(x)\varphi(y)$ для всех $x, y \in \Sigma^*$. Очевидно, произвольный морфизм полностью определяется образами символов, которые мы называем *блоками*.

Пусть для некоторого символа $a \in \Sigma$ блок $\varphi(a)$ начинается с a. В этом случае говорят, что φ допускает *неподвижную точку* $w = w(\varphi, a)$, т. е. бесконечное слово, начинающееся с a и удовлетворяющее равенству $w = \varphi(w)$. Неподвижная точка w может быть получена также как предел

$$w = \lim_{i \to \infty} \varphi^i(a).$$

Другим термином для обозначения неподвижной точки морфизма является термин *DOL-последовательность* (или *DOL-слово*).

DOL-последовательности — хорошо изученный и во многих отношениях полезный класс бесконечных слов. Опыт показывает, что часто

именно в этом классе целесообразно искать примеры последовательностей с заданными свойствами, например последовательностей, избегающих какого-либо слова над алфавитом переменных. Особенно часто в практике встречаются DOL-слова, обладающие сразу тремя полезными свойствами: равноблочностью, маркированностью и циркулярностью.

Морфизм φ называется *равноблочным*, если все его блоки имеют равную длину: $|\varphi(a_i)| = m$ при всех $i \in \{1, \dots, q\}$; неподвижная точка равноблочного морфизма называется *равноблочным* DOL-словом. Морфизм называется *маркированным*, если все его блоки начинаются с разных символов и кончаются разными символами:

$$\varphi(a_i) = a_{\pi(i)} s_i a_{\sigma(i)}$$

для каждого $i \in \{1, \dots, q\}$; здесь π, σ — перестановки над q элементами, а s_i — произвольные слова над Σ. Неподвижная точка маркированного морфизма называется *маркированным* DOL-словом.

Рассмотрим допустимое слово $u \in F$. Пара (u_1, u_2) называется *точкой синхронизации* слова u над w, если $u = u_1 u_2$ и между u_1 и u_2 в любом вхождении слова u в w проходит граница между блоками. Другими словами, рассмотрим все встречающиеся в w способы разбиения слова u на блоки; в точке синхронизации во всех блоках проходит граница между блоками (рис. 1).

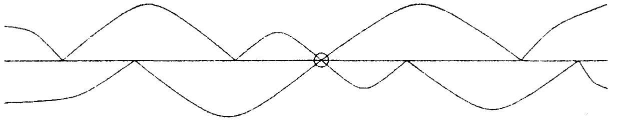

Рис. 1. Точка синхронизации

Слово из F называется *циркулярным*, если оно имеет по крайней мере одну точку синхронизации. Заметим, что если морфизм φ равноблочен *или* маркирован, то циркулярность слова u равносильна тому, что в u однозначно определяются *все* границы между блоками.

Говорят, что DOL-последовательность w *циркулярна* с *длиной синхронизации*, если любое допустимое в w слово длины не меньше L циркулярно. Определение циркулярности было введено в работах [8] и [2].

В данной статье устанавливается вид графов Рози для равноблочных маркированных циркулярных DOL-последовательностей. К этому классу принадлежит, например, слово Туэ–Морса — классический пример бескубной последовательности, являющийся неподвижной точкой морфизма φ_{TM}:

$$\begin{cases} \varphi_{TM}(a) = ab, \\ \varphi_{TM}(b) = ba. \end{cases}$$

Этот же класс содержит и такие известные примеры, как последовательность Дежан [4] и последовательность Керянена [7].

Предложение 1. *Язык подслов всякой маркированной DOL-последовательности продолжаем.*

Доказательство. Без ограничения общности можно предположить, что все символы алфавита в последовательности встречаются. Известно, что язык подслов DOL-последовательности продолжаем тогда и только тогда, когда первый символ последовательности встречается в ней еще где-то (см. теорему V.2 в [9]). Пусть последовательность начинается с символа a_1; тогда второй раз a_1 встречается в ней как последний символ в блоке $\varphi(a_{\sigma^{-1}(1)})$.

3. Определение графов Рози

Графом Рози́ порядка n последовательности w называется ориентированный граф $R(n) = (F(n), E(n))$, в котором множество вершин совпадает с множеством подслов длины n последовательности w, а множество дуг $E(n)$ соответствует ее подсловам длины $n+1$, т. е.

$$E(n) = \{(au, ub) \mid au, ub \in F(n), a, b \in \Sigma, aub \in F(n+1)\}.$$

Заметим, что $R(n)$ является подграфом графа де Брёйна порядка n.

Тип каждой вершины графа Рози совпадает с типом соответствующего подслова последовательности w; специальная вершина соответствует специальному слову, а ординарная вершина — ординарному слову.

Нетрудно понять, что маршруты длины l в графе Рози порядка n можно отождествить со словами длины $l+n$ над алфавитом Σ: маршрут, проходящий через вершины $v_1 \dots v_n$, $v_2 \dots v_{n+1}$, $\dots$, $v_{l+1} \dots v_{l+n}$, где $v_i \in \Sigma$, соответствует слову $v_1 v_2 \dots v_{l+n-1} v_{l+n}$. Далее будем обозначать маршруты (и, в частности, дуги, понимаемые как маршруты длины 1) в графах Рози именно таким образом. Заметим, что если все вершины маршрута, кроме, может быть, первой и последней, ординарны, то соответствующее слово всегда является подсловом последовательности w. В частности, дуги графа Рози $R(n)$ в точности соответствуют подсловам последовательности w длины $n+1$, т. е. вершинам графа $R(n+1)$.

4. Переход от $R(n)$ к $R(n+1)$

Как следует из замечания в конце предыдущего раздела, вершины графа $R(n+1)$ однозначно определяются графом $R(n)$. Более того, по графу $R(n)$ мы можем многое сказать и о дугах графа $R(n+1)$.

Действительно, рассмотрим вершину l_0vr_0 графа $R(n)$ и инцидентные ей дуги: как входящие из вершин вида l_il_0v, где $i = 1, \dots, l(l_0vr_0)$, так и выходящие в вершины вида vr_0r_j, где $j = 1, \dots r(l_ovr_0)$; здесь l_i и r_j — символы, а v — слово длины $n-2$. Входящие дуги в $R(n+1)$ соответствуют вершинам вида $l_il_0vr_0$, а выходящие — вершинам вида $l_0vr_0r_j$. Некоторые из этих вершин графа $R(n+1)$ соединены дугами, которые соответствуют подсловам последовательности w вида $l_il_0vr_0r_j$.

Заметим, что когда слово l_0vr_0 не является биспециальным, а язык подслов последовательности w продолжаем, выбор дуг такого вида однозначен: реализуются *все* возможные варианты.

Действительно, пусть $l(l_0vr_0) = 1$, т. е. l_0vr_0 не является специальным слева и может быть влево продолжено только одним символом l_1. Тогда в силу продолжаемости слово $l_1l_0vr_0$ может быть продолжено вправо теми же символами, что и l_0vr_0: в графе $R(n+1)$ оно соединено дугами со всеми словами $l_0vr_0r_j$. Аналогично если $r(l_0vr_0) = 1$, то в слово l_0vr_0 входят дуги из всех возможных слов $l_il_0vr_0$. Заметим, что в этих случаях слову длины n с данной (соответственно правой или левой) степенью специальности соответствует слово длины $n+1$ с той же степенью специальности.

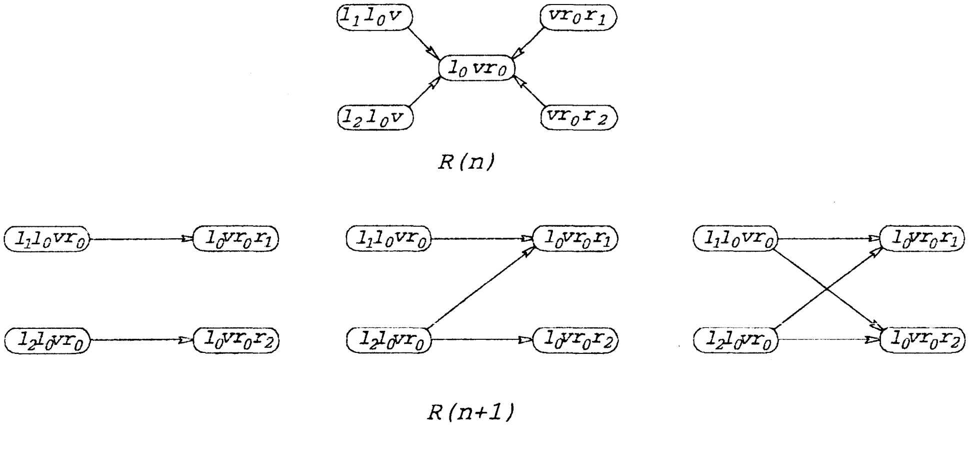

Рис. 2

Если же слово l_0vr_0 биспециально, мы не можем практически ничего утверждать о соответствующих дугах в $R(n+1)$. Необходима дополнительная информация о том, какие именно из слов вида $l_il_0vr_0r_j$ являются подсловами последовательности w. Например, на рис. 2 изображены три принципиально разные возможности, которые могут реализоваться для слова типа $(2, 2)$.

5. Гомеоморфизм графов Рози

Рассмотрим последовательность, первые разности комбинаторной сложности которой ограничены сверху константой, т. е. существует такое k, что для любого n

$$s(n) = f(n+1) - f(n) \leqslant k.$$

Очевидно, что сама комбинаторная сложность такой последовательности имеет линейный порядок роста: $f(n) \leqslant kn$. Ж. Кассень доказал [3], что верно и обратное: первые разности функции комбинаторной сложности, имеющей линейный порядок роста, ограничены сверху константой. Из этого следует, в частности,

Предложение 2. *Последовательность имеет линейный порядок роста комбинаторной сложности тогда и только тогда, когда ее графы Рози разбиваются на конечное число классов, в каждом из которых все графы гомеоморфны.*

Доказательство. Если комбинаторная сложность растет линейно, то ее первые разности ограничены. Далее, количество слов длины n, специальных справа, не может превышать первую разность $s(n)$: ведь не специальное справа слово длины n является префиксом только одного подслова последовательности длины $n+1$, а специальное справа слово является префиксом двух или более подслов. Точно так же не может превышать первую разность число слов данной длины, специальных слева. Таким образом, если первые разности ограничены, то ограничено константой и число специальных слева или справа слов любой заданной длины. Значит, ограничено сверху число вершин в схемах графов Рози любого порядка, а следовательно, число различных таких схем конечно. Поскольку каждой схеме соответствует свой класс гомеоморфности, из линейности комбинаторной сложности вытекает конечность числа классов гомеоморфности графов Рози.

Напротив, пусть число классов гомеоморфности графов Рози бесконечно. Тогда с возрастанием порядка графа Рози неограниченно возрастает максимальное количество вершин в его схеме, т. е. число левых или правых специальных слов данной длины. Следовательно, неограниченно возрастают и первые разности, а значит, из результата работы [3] следует, что комбинаторная сложность растет быстрее.

В [1] найдена формула для комбинаторной сложности равноблочной маркированной циркулярной DOL-последовательности. В частности, из этой формулы следовал линейный порядок роста функции комбинаторной сложности. Значит, графы Рози каждой такой последовательности

разбиваются на конечное число классов гомеоморфности. В следующем разделе опишем эти классы в явном виде.

6. Основной результат

Начиная с этого момента рассматриваем неподвижную точку w маркированного и равноблочного с длиной блока m морфизма φ. Полагаем также, что w циркулярна с длиной синхронизации L, и используем перестановки π и σ из определения маркированности морфизма.

Пусть K — наименьшее натуральное число, удовлетворяющее неравенству $m(K-1)+1 \geqslant L$.

Предложение 3. *Для каждого $n \geqslant K$ существует единственная тройка параметров разложения $(p(n), k(n), \Delta(n))$, где $p(n)$ — целое неотрицательное число, $k(n) \in \{K, \dots, m(K-1)\}$, $\Delta(n) \in \{1, \dots, m^{p(n)}\}$ и*

$$n = m^{p(n)}(k(n)-1) + \Delta(n).$$

Доказательство проведем индукцией по n. Нетрудно видеть, что равенству $n = K$ соответствует тройка $(0, K, 1)$. Пусть $n \geqslant K$ и параметры разложения числа n равны (p, k, Δ). Для параметров разложения числа $n+1$ возможны три случая:

- если $\Delta \neq m^p$, то они равны $(p, k, \Delta+1)$;
- если $\Delta = m^p$, но $k \neq m(K-1)$, то они равны $(p, k+1, 1)$;
- наконец, если $\Delta = m^p$ и $k = m(K-1)$, то они равны $(p+1, K, 1)$.

Ясно, что в любом из трех случаев параметры разложения числа $n+1$ определены корректно и однозначно.

Заметим, что при неограниченном росте n параметр $k(n)$ принимает лишь конечное число значений — от K до $m(K-1)$. На этом факте основаны формулы для комбинаторной сложности и частоты слов последовательности w, полученные соответственно в [5] и [6]: сначала вычисляются значения этих функций для всех возможных значений $k(n)$, затем эти значения подставляются в общую формулу в качестве коэффициентов. Следующая теорема дает результат такого же рода для графов Рози последовательности w.

Теорема 1. *Для любого $n \geqslant K$*
если $\Delta(n) = m^{p(n)}$, то $R(n) \simeq R(k(n))$;
если $\Delta(n) \neq m^{p(n)}$, то $R(n) \simeq R(mk(n)-1)$.

Таким образом, найдя графы Рози для нескольких начальных длин подслов, мы узнаем, как — с точностью до гомеоморфизма — выглядят *все* графы Рози данной последовательности.

7. Доказательство теоремы 1

Доказательство теоремы 1 разбивается на несколько утверждений.

Предложение 4. *Пусть v — слово длины не менее K. Тогда $l(v) = l(\varphi(v))$ и $r(v) = r(\varphi(v))$.*

Доказательство. Пусть v может быть продолжено вправо до подслова слова w буквами $a_{i_1}, \ldots, a_{i_l}$, а влево буквами $a_{j_1}, \ldots, a_{j_r}$. Тогда слово $\varphi(v)$ может быть продолжено вправо буквами $a_{\pi(i_1)}, \ldots, a_{\pi(i_l)}$, а влево буквами $a_{\sigma(j_1)}, \ldots, a_{\sigma(j_r)}$. Так как в силу циркулярности слово $\varphi(v)$ однозначно разбивается на блоки, то нет других букв, которыми оно может быть продолжено вправо или влево. Поэтому $l(v) = l(\varphi(v)) = l$ и $r(v) = r(\varphi(v)) = r$.

Предложение 5. *Всякое циркулярное биспециальное слово $u \in F$ является образом другого биспециального слова v, т. е. $u = \varphi(v)$.*

Доказательство. Из определений циркулярности и маркированности следует, что во всяком циркулярном слове однозначно определены все границы между блоками. Более того, если слово начинается с неполного блока, то в силу маркированности этот блок однозначно восстанавливается до полного. Значит, специальным слева может быть только слово, начинающееся с целого блока. Аналогично специальным справа может быть только слово, кончающееся полным блоком. Значит, биспециальным может быть только слово, состоящее из полных блоков, т. е. являющееся образом другого слова. В силу предыдущего предложения прообраз биспециального слова биспециален.

Предложение 6. *Подслово длины $n \geqslant K$ последовательности w может быть биспециальным только при $\Delta(n) = m^{p(n)}$.*

Доказательство проводится индукцией по $p(n)$. Пусть $p(n) = 0$. Тогда $\Delta(n)$ всегда равно $1 = 2^0$ и утверждение справедливо. Пусть $p(n) > 0$. Тогда каждое слово u длины n циркулярно. Если u биспециально, то в силу предыдущего предложения $n = |u| = |\varphi(v)| = m|v|$, где v — некоторое биспециальное слово. Заметим, что $p(|v|) = p(n) - 1$, а $k(|v|) = k(n)$. Значит, по предположению индукции $|v| = m^{p(n)-1}k(n)$. Следовательно, $n = m|v| = m^{p(n)}k(n)$.

Предложение 7. *Для всех $n \geqslant K$*

$$R(n) \simeq R(mn).$$

Доказательство. Из рассуждений, использованных в доказательстве предложения 5, следует, что каждое специальное (справа или слева) слово длины mn является образом какого-либо слова длины n. Поэтому

в силу предложения 4 отображение φ задает изоморфизм между множествами специальных вершин графов $R(n)$ и $R(mn)$. Более того, если u — прямой маршрут, соединяющий специальные вершины s_1 и s_2 графа $R(n)$ (напомним, что в разд. 3 мы договорились обозначать маршруты в графах Рози словами), то вершины $\varphi(s_1)$ и $\varphi(s_2)$ соединены в $R(mn)$ маршрутом $\varphi(u)$. Нетрудно видеть, что и этот маршрут прямой.

Заметим, что других прямых маршрутов в графе $R(mn)$ нет. Ведь число исходящих из специальной вершины прямых маршрутов равно ее полустепени исхода, т. е. правой степени специальности соответствующего слова. Значит, в силу предложения 4 число прямых маршрутов, равное сумме полустепеней исхода всех специальных вершин, одно и то же в $R(n)$ и $R(mn)$, и соответствие между прямыми маршрутами задается отображением φ.

Мы установили, что φ задает изоморфизм схем графов $R(n)$ и $R(mn)$, а значит, и гомеоморфизм самих этих графов.

Предложение 8. *Если $n \geqslant K$ и $\Delta(n) \neq m^{p(n)}$, то*

$$R(n) \simeq R(m^{p(n)}k(n) - 1).$$

Доказательство. Поскольку согласно предложению 8 биспециальных слов длины n не существует, а язык F продолжаем, то граф $R(n+1)$ однозначно восстанавливается по графу $R(n)$ так, как описано в разд. 4. При этом каждому специальному справа или слева слову длины n путем приписывания соответственно слева или справа символа соответствует единственное специальное слово длины $n+1$. Если $\Delta(n) \neq m^{p(n)} - 1$, то биспециальных слов длины $n+1$ также не существует, и мы получаем взаимно однозначное соответствие между специальными словами длины n и длины $n+1$. Нетрудно проверить, что соответствие существует и между прямыми маршрутами в $R(n)$ и $R(n+1)$, т. е. что при $\Delta(n) \neq m^{p(n)} - 1$ графы $R(n)$ и $R(n+1)$ гомеоморфны. Следовательно, все графы Рози для длин из интервала $[m^{p(n)}(k(n) - 1),\ m^{p(n)}k(n) - 1$ гомеоморфны друг другу и, в частности, графу $R(m^{p(n)}k(n) - 1)$.

Доказательство теоремы 1. Для доказательства теоремы первого утверждения достаточно многократно применить предложение 7: если $\Delta(n) = m^{p(n)}$, т. е. если $n = m^{p(n)}k(n)$, то

$$R(n) = R(m^{p(n)}k(n)) \simeq R(m^{p(n)-1}k(n)) \simeq \ldots \simeq R(mk(n)) \simeq R(k(n)).$$

Для доказательства второго утверждения теоремы воспользуемся предложениями 7 и 8. В результате получаем

$$R(m^{p(n)}k(n) - 1) \simeq R(m^{p(n)}k(n) - m) \simeq R(m^{p(n)-1}k(n) - 1).$$

Следовательно,

$$R(m^{p(n)}k(n)-1) \simeq R(m^{p(n)-1}k(n)-1) \simeq \ldots \simeq R(mk(n)-1).$$

Еще раз воспользовавшись предложением 8, получаем

$$R(n) \simeq R(m^{p(n)}k(n)-1) \simeq R(mk(n)-1).$$

8. Графы Рози последовательности Туэ–Морса

В этом разделе продемонстрируем применение теоремы 1, охарактеризовав графы Рози последовательности Туэ–Морса, определение которой было дано в разд. 2. Эта последовательность равноблочна с длиной блока $m = 2$ и маркирована; известно, что она циркулярна с длиной синхронизации $L = 4$. Нетрудно видеть, что $K = 3$. Поэтому при любом n параметр $k(n)$ равен либо 3, либо 4. Таким образом, в этом случае классы гомеоморфности графов Рози задаются графами $R(3)$, $R(4)$, $R(5)$ и $R(7)$. Их же нетрудно построить вручную (рис. 3).

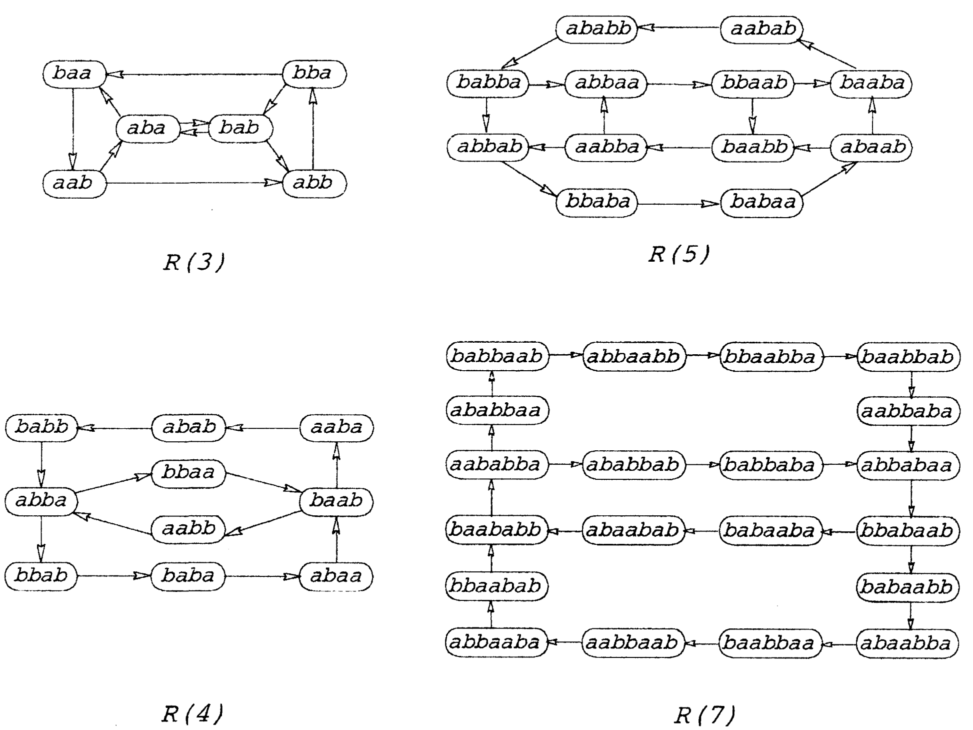

Рис. 3

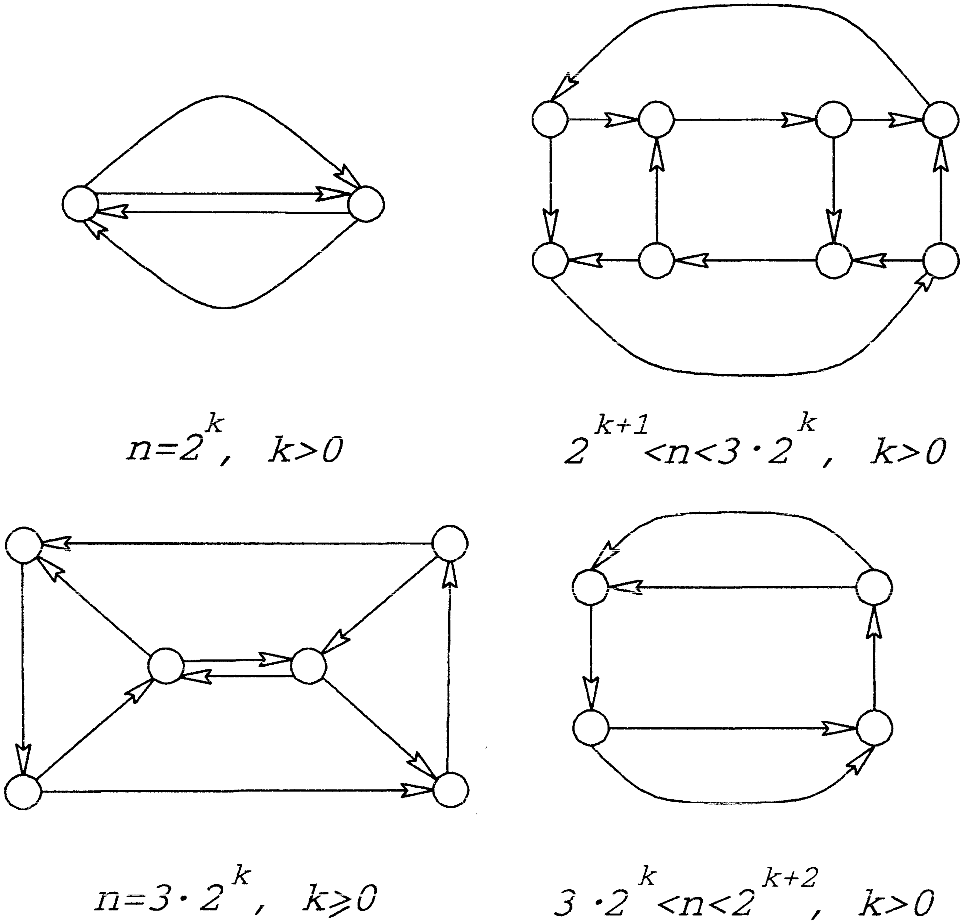

Рис. 4

Перейдя теперь от графов с рис. 2 к их схемам и заметив, что $R(2) \simeq R(4)$, получим характеризацию графов Рози, изображенную на рис. 4.

ЛИТЕРАТУРА

1. **Фрид А. Э.** О комбинаторной сложности итеративно порождаемых символьных последовательностей // Дискрет. анализ и исслед. операций. Сер. 1. 1997. Т. 4, № 1. С. 53–59.
2. **Cassaigne J.** An algorithm to test if a given circular HD0L-language avoids a pattern // Information processing'94. V. 1. Amsterdam: North-Holland, 1994. P. 459–464.

3. **Cassaigne J.** Special factors of sequences with linear subword complexity // Developments in Language Theory II. River Edge, NJ: World Sci. Publ., 1996. P. 25–34.
4. **Dejean F.** Sur un théorème de Thue // J. Combin. Theory. Ser. A. 1972. V. 13, N 1. P. 90–99.
5. **Frid A.** On uniform D0L words // Theoretical Aspects of Computer Science. Berlin: Springer, 1998. P. 544–554. (Lecture Notes in Comput. Sci.; V. 1373).
6. **Frid A.** On the frequency of factors in a D0L word // J. of Automata, Languages and Combinatorics. 1998. V. 3, N 1. P. 29–41.
7. **Keränen V.** Abelian squares are avoidable on 4 letters // Automata, Languages and Programming. Berlin: Springer, 1992. P. 41–52. (Lecture Notes in Comput. Sci.; V. 623).
8. **Mignosi F., Séébold P.** If a D0L language is k-power free then it is circular // Automata, Languages and Programming. Berlin: Springer, 1993. P. 507–518. (Lecture Notes in Comput. Sci.; V. 700).
9. **Queffelec M.** Substitution dynamical systems — spectral analysis. Berlin: Springer-Verl., 1987. (Lecture Notes in Math.; V. 1294).
10. **Rauzy G.** Suites à termes dans un alphabet fini // Seminar on Number Theory, 1982–1983. Talence: Univ. Bordeaux I, 1983. Exp. N 25. 16 p.

Адрес автора:

Институт математики
им. С. Л. Соболева СО РАН,
пр. Академика Коптюга, 4,
630090 Новосибирск, Россия

Статья поступила
11 марта 1999 г.