Доклады Академии наук СССР
1971. Том 201, № 4

УДК 517.11

МАТЕМАТИКА

Я. С. СМЕТАНИЧ

О ВОССТАНОВЛЕНИИ СЛОВ

(Представлено академиком П. С. Новиковым 12 V 1971)

Пусть некоторым известным способом слову X сопоставлена совокупность его перекрывающихся подслов; можно ли, зная эту совокупность, однозначно и достаточно эффективным способом восстановить слово X — такова нестрогая постановка задачи восстановления слова. Математическая задача восстановления слова возникает при анализе методов определения первичной структуры гетерополимеров (t-РНК, белок). Макромолекула полимера с точки зрения первичной структуры есть слово в некотором алфавите. Один из способов определения первичной структуры заключается в получении перекрывающихся фрагментов молекулы, по которым восстанавливается первичная структура целой молекулы. Слова, соответствующие, например, молекуле ДНК, могут иметь длину порядка 10^6. В п. 1 заметки мы определяем специальный класс функций, регулярных функций. Класс регулярных функций охватывает, по-видимому, существующие способы «расщепления» молекулы на фрагменты (([1, 2]) и др.). В п. 2 дается точная постановка задачи восстановления слова и формулируются результаты.

1. Мы будем рассматривать слова в конечном алфавите Σ и обозначать их большими латинскими буквами. Однобуквенные слова в Σ обозначаются малыми латинскими буквами, $[X$, Δ означают соответственно длину слова X и пустое слово; равенство $X = Y$ слов X и Y понимается в смысле их графического тождества. Добавим к Σ знак $*$ и обозначим расширенный алфавит Σ'. Пусть слово X представлено в виде $X = X_1AX_2$, где $A \neq \Delta$. Слово $X_1 * A * X_2$ в Σ' назовем о т р е з к о м с л о в а X, слово A — р е а л и з а ц и е й этого отрезка. Если $X_1 = \Delta$, то получаем начальный отрезок слова X, если $X_2 = \Delta$,— концевой. Начальные и концевые отрезки будем называть граничными отрезками. Отрезки будем обозначать большими латинскими буквами с чертой сверху.

Пусть $X = X_1B_1AB_2X_2$, где $A \neq \Delta$ и $B_1B_2 \neq \Delta$. Рассмотрим два отрезка слова X: $X_1 * B_1AB_2 * X_2$ и $X_1B_1 * A * B_2X_2$, обозначим их соответственно через $\bar{C}_1$ и $\bar{C}_2$. Будем говорить, что отрезок $\bar{C}_2$ п о к р ы в а е т с я отрезком $\bar{C}_1$ в слое X. Каждое слово X порождает множество всех своих отрезков $V(X)$. Рассмотрим произвольное подмножество $\alpha(X)$ множества $V(X)$. Множество $\alpha(X)$ назовем п о к р ы т и е м, если каждый отрезок слова X, реализация которого есть однобуквенное слово, покрывается некоторым отрезком из $\alpha(X)$.

Пусть слово X представлено в виде $X = X_1AX_2 = Y_1BY_2$. Будем писать

$$X_1 * A * X_2 \prec Y_1 * B * Y_2,$$

если $[X_1 < [Y_1$ и $[X_2 > [Y_2$.

Назовем отрезки $\bar{A}$, $\bar{B}$ слова X п е р е к р ы в а ю щ и м и с я, если

1) $X = X_1C_1aC_2bC_3X_3$, где $C_2 \neq \Delta$;
2) $\bar{A} = X_1 * C_1aC_2 * bC_3X_3$;
3) $\bar{B} = X_1C_1a * C_2bC_3 * X_3$.

При этом будем говорить, что $\bar{A}$ справа перекрывается с $\bar{B}$, а $\bar{B}$ слева перекрывается с $\bar{A}$ в слове X.

Покрытие $\alpha(X)$ назовем **простым**, если не существует $\bar{A}$, $\bar{B} \in \alpha(X)$, таких, что $\bar{A}$ покрывается $\bar{B}$ в X.

Простое покрытие $\alpha(X)$ назовем **связным**, если для каждого неконцевого отрезка $\bar{A} \in \alpha(X)$ найдется отрезок $\bar{B} \in \alpha(X)$ такой, что $\bar{A}$ справа перекрывается с $\bar{B}$ и для каждого неначального отрезка $\bar{A} \in \alpha(X)$ найдется отрезок $\bar{B} \in \alpha(X)$ такой, что $\bar{A}$ слева перекрывается с $\bar{B}$.

Рассмотрим связное покрытие $\alpha(X)$. Легко видеть, что отношение $\prec$ упорядочивает все отрезки из $\alpha(X)$, при этом первый в этом упорядочении элемент $\alpha(X)$ является единственным в $\alpha(X)$ начальным отрезком, а последний — единственным концевым. Перекрывающиеся отрезки $\bar{A}$, $\bar{B} \in \alpha(X)$, где $\bar{A} \prec \bar{B}$ назовем **соседями**, если не существует $\bar{C} \in \alpha(X)$ такого, что $\bar{A} \prec \bar{C} \prec \bar{A}$. Пусть $\bar{A}$ — неконцевой отрезок из $\alpha(X)$ и пусть $\bar{B}$ — его (очевидно, единственный) правый сосед; пусть, наконец, (1), (2), (3) — соответствующие представления слова X и отрезков $\bar{A}$, $\bar{B}$. Отрезок $C_1 * a * C_2$ слова C_1aC_2 назовем **левой отметкой** отрезка $\bar{A}$ (реализации $\bar{A}$). Аналогично определяется понятие **правой отметки** для каждого неначального отрезка из $\alpha(X)$. Таким образом, в каждом отрезке $\alpha(X)$, кроме граничных, выделяются два однобуквенных отрезка: левая отметка и правая отметка. В начальном отрезке выделяется только левая отметка, в концевом — только правая отметка.

Рассмотрим связное покрытие $\alpha(X)$ с выделенными отметками во всех его отрезках. Для отрезков из $\alpha(X)$, отличных от граничных, введем отношение ∞. $\bar{A} \infty \bar{B}$ имеет место тогда и только тогда, когда выполняются два условия:

a) реализация $\bar{A}$ = реализация $\bar{B}$;

b) левая и правая отметки $\bar{A}$ совпадают соответственно с левой и правой отметками $\bar{B}$.

Отношение ∞ разбивает все отрезки из $\alpha(X)$, кроме граничных, на классы эквивалентности. Для каждого такого класса K образует упорядоченную четверку $\langle A, \bar{a}, \bar{b}, k\rangle$, где A — реализация отрезков из K, $\bar{a}$ и $\bar{b}$ — левая и правая отметки A, k — число элементов K. Для начального и концевого отрезков образуем соответственно четверки $\langle A_0, \bar{a}, \Delta, 1\rangle$ и $\langle B_0, \Delta, \bar{b}, 1\rangle$, где A_0 и B_0 — реализации граничных отрезков, $\bar{a}$ и $\bar{b}$ — левая и правая отметки граничных отрезков. Объединим все четверки, соответствующие классам K и граничным отрезкам в одно множество, которое обозначим $M[\alpha(X)]$.

Поставим в соответствие каждому слову X некоторое его связное покрытие $\alpha(X)$. Получим функцию $\alpha(X)$. Каждая функция $\alpha(X)$ порождает функцию $M[\alpha(X)]$, которую назовем **нормальной**.

Нормальную функцию $h(X)$ назовем **регулярной**, если для любой пары слов X и Y и для каждой нормальной функции $g(X)$ из равенства $g(Y) = h(X)$ следует равенство $h(X) = h(Y)$.

2. Для класса регулярных функций сформулируем задачу восстановления слова. По известному значению произвольной регулярной функции $h(X)$ при неизвестном слове X построить слово Y такое, что $h(X) = h(Y)$ и определить, существует или нет слово Z, $Z \neq Y$, такое, что $h(X) = h(Z)$. Если слова Z не существует, то слово X **восстановимо** по значению $h(X)$.

Сформулированная алгоритмическая задача решается, например, перебором всех слов некоторой фиксированной длины. Поэтому она требует следующего уточнения: найти алгоритм, решающий задачу, с оценкой числа шагов.

Переходим к формулировке результатов. **Длиной отрезка** назовем длину его реализации. Пусть $\beta(X)$ — связное покрытие слова X всеми его отрезками, длина которых равна 2. Легко видеть, что $s(X) = M[\beta(X)]$ есть регулярная функция. Значение $s(X)$ назовем **составом второго ранга слова** X. Пусть слово X можно представить в одной из форм

$$X = X_1aPbX_2aQbX_3, \qquad X = Z_1dRdSdZ_2.$$

Тогда каждое из слов

$$Y_1 = X_1aQbX_2aPbX_3, \qquad Y_2 = Z_1dSdRdZ_2$$

есть результат регулярного преобразования слова X.

Теорема (о составе второго ранга). *Для того чтобы слова X и Y имели один и тот же состав второго ранга, т. е. чтобы выполнялось равенство $s(X) = s(Y)$, необходимо и достаточно, чтобы существовала последовательность слов $X_1, X_2, \ldots, X_m$ такая, что $X = X_1$, $Y = X_m$ и X_{k+1} есть результат регулярного преобразования X_k $(1 \leqslant k \leqslant m-1)$.*

Эта теорема имеет следующее обобщение: *пусть $h(X)$ — произвольная регулярная функция, пусть $h(X_1) = h(X_2)$.*

Тогда слово X_1 переводится в слово X_2 некоторыми преобразованиями, аналогичными регулярным.

С помощью этого обобщения можно показать, что задача восстановления слова для класса всех регулярных функций сводится к задаче восстановления слова по составу второго ранга. На основании этого сведения и теоремы о составе второго ранга строится алгоритм, решающий задачу восстановления слова для класса всех регулярных функций, число элементарных операций которого имеет порядок n^4, где n — длина восстанавливаемого слова X, и за элементарную операцию принято сравнение двух однобуквенных слов.

Приведем в заключение два примера связных покрытий, которые порождают регулярные функции.

1) Естественным обобщением состава второго ранга является покрытие $\beta_k(X)$ слова X всеми его отрезками, длина которых есть k $(k > 2)$.

2) Второй пример построим неформально. Пусть выделены все вхождения буквы a в слове X. «Разрезая» X непосредственно после каждого из вхождений буквы a, получим множество отрезков M_a, имеющих концевой единственную букву a (кроме, возможно, концевого отрезка X). Аналогично строим множество M_b с помощью всех вхождений b в X. Объединяя M_a и M_b и выбросив отрезки суммы, которые покрываются другими отрезками суммы, получим связное покрытие.

Институт биологической физики
Академии наук СССР
Пущино-на-Оке

Поступило
10 V 1971

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] В. Г. Туманян, Л. Л. Киселев, Биофизика, 8, 147 (1963). [2] Р. Холли, Молекулы и клетки, в 3, М., 1968, стр. 77.