

УДК 519.766

НОВЫЕ АСПЕКТЫ МЕТОДА ИНДУКТИВНОЙ РЕКОНСТРУКЦИИ
ГРАММАТИК

М.К. Тимофеева

I. В [1] описан метод индуктивной реконструкции грамматик, позволяющий осуществлять переход вида $\langle T, S \rangle \rightarrow G$, где T - текст, S - система априорных знаний о языке, породившем текст T , G - формальное описание этого языка. Этот метод, как он реализован в [1]; предполагает наличие в тексте определенных выделенных символов-разделителей. В рассматривавшихся текстах естественных языков это - пробелы и знаки пунктуации. В других текстах не всегда можно выделить такие символы. Например, в генетических текстах - упорядоченных последовательностях мономеров (нуклеотидов или аминокислот) - роль знаков пунктуации выполняют не одиночные символы, а их последовательности [2].

В данной статье предлагается модификация метода [1], адаптированная к анализу текста без разделителей. Эта модификация позволила расширить возможности метода [1] и в другом отношении: перейти от выявления грамматической системы, задающей линейные характеристики подцепочек текста, к выявлению грамматической системы, задающей нелинейную структуру таких цепочек и отражающую взаимообусловленность подцепочек текста, расположенных на расстоянии друг от друга.

Прежде чем анализировать тексты неизвестных языков, следует проверить качество работы предлагаемого варианта метода реконструкции на текстах какого-либо известного языка, например, естественного. Эксперименты, проведенные на текстах русского языка, показали, что грамматические структуры, выявляемые на этих текстах, являются осмысленными с позиции традиционной грамматики русского языка.

2. Система S априорных знаний о языке, породившем анализируемый текст, формируется из заданного множества элементарных предположений. Это множество является открытым, т.е. может по мере необходимости пополняться новыми предположениями. В данной статье предлагается расширение системы элементарных предположений, сущность которого можно пояснить следующим образом.

В [1] объектами исследования были только цепочки, представляющие собой осмысленное целое с позиции рассматриваемого языка (при анализе текста естественного языка это – словоформы – отрезки текста от пробела до пробела). В таком случае: а) анализируемая цепочка всегда состоит из целого числа грамматических единиц; б) всегда ясно, какую цепочку можно выбрать в качестве объекта анализа и какие вхождения этой цепочки в текст следует при таком анализе учитывать.

В предлагаемой модификации метода [1] объектом анализа может быть произвольная подцепочка $w = a_1 \dots a_k$ заданного текста (a_1, a_2, \dots, a_k – символы алфавита). Такая подцепочка может не представлять собой осмысленного целого. В частности, w может состоять из нецелого числа языковых единиц. В этом случае предлагается анализировать ее подцепочки с "плавающими" левой и правой границами (последовательность анализируемых подцепочек w_i^j иллюстрируется рис.1). Принципы анализа каждой цепочки w_i^j аналогичны использованным в [1]. Дополнение системы S состоит во введении предположений об интерпретации соотношений между результатами анализа всех цепочек w_i^j ($1 \leq i \leq k-1, 1 \leq j \leq k, i < j$) для заданной цепочки w .

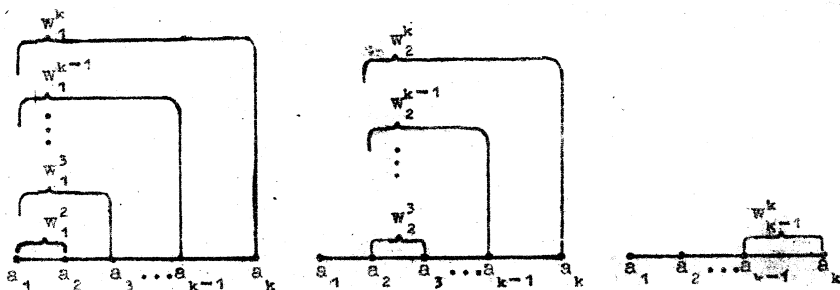


Рис. 1

Предлагаемое расширение системы S несколько меняет сам принцип анализа текста. Метод [I] позволяет выявить грамматику, задающую правила комбинирования грамматических единиц в заданном тексте. Это - характеристика текста в целом, не отражающая процесса его восприятия. Реально текст воспринимается не в целом, а читается последовательно символ за символом. Прочитанная часть текста служит основанием для некоторого прогноза относительно последующей его части. Использование "плавающих" границ анализируемой цепочки позволяет проследить изменение ожидаемой структуры при продвижении по тексту слева направо и выявить влияние каждого символа заданной цепочки на структуру остальной части этой цепочки.

3. Для формального описания предлагаемой в статье модификации метода [I] введем некоторые обозначения и определения.

Условные обозначения:

A - алфавит;

A* - множество всех цепочек в алфавите A;

Λ - пустая цепочка;

T - заданный текст в алфавите A;

w - цепочка, структура которой исследуется, w - подцепочка T,

$w = a_1 \dots a_k, a_i \in A, 1 \leq i \leq k$;

$n_i(w)$ - частота встречаемости в тексте T цепочки, совпадающей с первыми i символами цепочки w;

$w_w^s = \{w_i = a_1 a_{i+1} \dots a_k \mid 1 \leq i \leq s\}$, где $0 < s < k$.

Определения:

Символ a_i цепочки w называется неинформативным, если $n_{i-1}(w) = n_i(w)$, т.е. за цепочкой $a_1 \dots a_{i-1}$ в тексте T всегда следует символ a_i . Символ a_i цепочки w называется информативным, если $n_{i-1}(w) \neq n_i(w)$, т.е. за цепочкой $a_1 \dots a_{i-1}$ в тексте T не всегда следует символ a_i .

Сегментация цепочки w есть цепочка $I_w = \sigma_1 \varepsilon_1 \sigma_2 \varepsilon_2 \dots \sigma_f \varepsilon_f$,

где

$$\sigma_i = \begin{cases} "+", & \text{если } \varepsilon_i - \text{информативные символы } w; \\ "-", & \text{если } \varepsilon_i - \text{неинформативные символы } w; \end{cases}$$

$w = \varepsilon_0 \varepsilon_1 \varepsilon_2 \dots \varepsilon_f, \sigma_i \neq \sigma_{i+1} (1 \leq i < f), \sigma_1 = "-",$ цепочка ε_0 состоит только из информативных символов цепочки w. Цепочка ε_0 будет называться исходной цепочкой для сегментации I_w ; цепочки ε_i , для которых $\sigma_i = "+",$ - информативными зонами; цепочки ε_i , для которых $\sigma_i = "-",$ - неинформативными зонами.

ПРИМЕР I. Рассмотрим цепочку \tilde{w} из текста $T = \text{INF}(\text{текст научно-технических рефератов длиной } \sim 3500 \text{ символов})$:

$\tilde{w} = \text{ВЫЧИСЛИТЕЛЬНЫЕ ЦЕНТРЫ КОЛЛЕКТИВНОГО ПОЛЬЗОВАНИЯ.}$

Сегментация цепочки \tilde{w} имеет вид:

$I_{\tilde{w}} = \text{И С Л И Т Е Л Ъ Н Ы Е Ц Е Н Т Р Ы К О Л Л Е К Т И В Н О Г О П О Л Ъ З О В А Н И Я,}$

исходная цепочка $g_0 = \text{ВЫЧ.}$

Цепочка w не сегментируема, если $I_w = \Lambda$. Как видно из предыдущего примера, цепочка "ВЫЧ" не сегментируема.

Построим множество цепочек $w_w^s = \{w_1, \dots, w_s\}$, содержащее только сегментируемые цепочки, такое, что цепочка $a_{s+1}a_{s+2} \dots a_k = w_{s+1}$ (состоящая из последних $k-s$ символов цепочки w) не сегментируема. Каждой цепочке $w_i \in w_w^s$ соответствует своя сегментация I_{w_i} со своей исходной цепочкой v_i .

Обозначим через J_w множество сегментаций цепочек из w_w^s . При выявлении структуры цепочки w элементы множества J_w рассматриваются в порядке возрастания индекса i , т.е. сначала I_{w_1} , затем I_{w_2} и т.д. В приложении приведен пример автоматического построения множества $J_{\tilde{w}}$ для цепочки \tilde{w} из примера I. Каждая строка машинной распечатки содержит один элемент из $J_{\tilde{w}}$.

Рассмотрим некоторую сегментацию $I_w = \sigma_1 g_1 \dots \sigma_t g_t \in J_w$. Устойчивость информативной зоны $g_i = a_{j_1} \dots a_{j_V}$ ($a_j \in \Lambda$, $j_1 \leq j_2 \leq j_V$) есть число элементов множества J_w , в которых цепочка $a_{j_1} \dots a_{j_V}$ является информативной зоной. Например, устойчивость информативной зоны $g_1 = a_{13} = \text{"И"}$ цепочки \tilde{w} равна 8 (см. приложение).

Пусть $I_{w_1}, I_{w_{i+1}}$ — две сегментации из J_w . Если элемент a_j содержится в обеих сегментациях и в $I_{w_{i+1}}$ он информативен, а в I_{w_1} — нет, то будем говорить, что a_i влияет на a_j (например, элемент $a_{14} = \text{"Е"}$ цепочки \tilde{w} влияет на элемент $a_{20} = \text{"И"}$ той же цепочки).

4. Предлагаемый в статье метод анализа текста состоит во введении следующих элементарных предположений о языке, породившем этот текст.

I. Граница, разделяющая структурные единицы цепочки w , проходит непосредственно слева от элемента a_i ($i > 2$), если в J_w имеется сегментация I_{w_j} такая, что a_i в ней информативен, а a_{i-1} — нет.

II. Граница, разделяющая структурные единицы цепочки w , проходит непосредственно справа от элемента a_i ($i > 1$), если в сегментации $I_{w_{i+1}}$ существует информативная зона, все элементы которой в сегментации I_{w_i} являются неинформативными.

III. Структурная единица e_q связана со структурной единицей e_q цепочки w , если некоторый символ цепочки e_q влияет на некоторый символ цепочки e_{q+1} .

ПРИМЕР 2. Цепочка \tilde{w} из примера I может быть разделена на структурные единицы следующим образом:

ВЫ/ЧИСЛ/ИТЕЛЬН/ЬЕ/ЦЕНТР/Ы/КОЛЛЕКТИВНО/ГО/ПОЛЬЗОВ/А/НИ/Я.

e_1 e_2 e_3 e_4 e_5 e_6 e_7 e_8 e_9 e_{10} e_{11} e_{12}

Границы между e_1 и e_2 , e_6 и e_7 выделены по предположению II; границы между e_2 и e_3 , e_3 и e_4 , e_5 и e_6 , e_7 и e_8 , e_9 и e_{10} , e_{10} и e_{11} , e_{11} и e_{12} выделены по предположению I; границы между e_4 и e_5 , e_8 и e_9 могут быть выделены по обоим указанным предположениям.

ПРИМЕР 3. Единица $e_4 = \text{"ЬЕ"}$, выделенная в цепочке \tilde{w} , связана с единицей $e_6 = \text{"И"}$, e_6 , в свою очередь, связана с единицей $e_8 = \text{"ГО"}$, а e_8 - с единицей $e_{12} = \text{"Я"}$, т.е. все окончания приведенного словосочетания оказываются последовательно связанными. На распечатке связи между структурными единицами обозначены стрелками.

Заданной цепочке w сопоставляется структура вида $Z_w = \langle E_w, P_w \rangle$, где E_w - набор структурных единиц w , выделенных по предположению I или II, P_w - отношение связанности между элементами w , установленное на основании предположения III.

Предположения I-III включены в систему элементарных предположений, описанную в [1]. При нахождении информативных и неинформативных зон использовались полные спектры 1-грамм [3] - списки частот встречаемости в заданном тексте различных его подцепочек длиной в 1 символ. Вместо полных спектров 1-грамм могут использоваться частичные спектры, программы получения которых описаны в [4].

5. В приложении приведены примеры машинных распечаток, полученные на тексте научно-технических рефератов INF (анализ текста слева направо и справа налево) и общественно-политическом тексте POL русского языка (~ 4000 символов), а также, для сравнения, - на генетическом тексте SV40 (~ 5000 символов).

По приведенным распечаткам видно, что границы разделения единиц в естественно-языковых текстах почти всегда имеют лингвистическое основание и соответствуют границам между морфемами. Ошибки в проведении таких границ, как правило, объясняются опечатками в тексте. Таковы, например, разделение

"ВЫЧИСЛИТЕЛЬНЫЕ ЦЕНТРЫ КОЛЛЕКТИВНОГО ПОЛЬЗОВАНИЯ",

полученное при анализе цепочки \tilde{w} справа налево, и разделение

"ПРОЛЕТАРСКОГО ИНТЕРНАЦИОНАЛИЗМА",

полученное при анализе подцепочки

δ = ПРОЛЕТАРСКОГО ИНТЕРНАЦИОНАЛИЗМА

текста POL слева направо (здесь указаны не все разделения цепочек \tilde{w} и δ). Выявленные связи между единицами также отражают действительные отношения, свойственные данному тексту. Они связывают как морфемы внутри слов (например, ВЫ → ИТЕЛЬН), так и морфемы разных слов (например, цепочка согласующихся окончаний НЕ → Н → ГО → Я).

Если проводить чисто формальное сопоставление результатов анализа естественно-языковых и генетических текстов (не учитывая семантики этих текстов), то можно обнаружить следующие различия:

Сегментации естественно-языковых текстов сложнее. Как правило, они содержат несколько информативных зон. Сегментации генетических текстов обычно имеют более простое строение вида $I_w = -g_1$ (в g_1 все символы неинформативны). Для естественно-языковых текстов характерны большая устойчивость информативных зон и большие длины исходных цепочек.

Применение описанного в статье метода реконструкции к анализу слитного текста позволило выявить ряд сложностей:

1) выбор способа совмещения результатов анализа текста слева направо и справа налево (как видно из приведенного в приложении примера, эти результаты в общем случае расходятся);

2) нахождение способа отбора анализируемых подцепочек при необходимости построения структуры всего текста (а не отдельных его подцепочек, как было описано выше);

3) создание программных средств, обеспечивающих возможность работы с текстами, спектры 1-грамм которых не помещаются целиком в оперативную память.

6. Помимо задачи анализа слитного текста неизвестного языка, описанный метод может служить средством решения ряда других задач.

Каждая группа информативных символов важна для восприятия всей анализируемой цепочки. По последовательности информативных зон можно однозначно воссоставить всю оставшуюся часть цепочки. Это свойство может быть использовано при выборе экономной кодировки или при восстановлении испорченных частей подцепочек текста. Изменение информативных и неинформативных зон при последовательном (по мере возрастания i) анализе сегментаций из J_w позволяет строить структуры, отражающие динамику восприятия цепочки w (на рис.2 приведена часть структуры, определяющей динамику восприятия цепочки \tilde{w} в тексте INF). Стрелка вида:

$$\boxed{a_j a_{j+1} \dots a_{j+p}} \rightarrow \boxed{a_{j+m} \dots a_{j+m+r}}$$

на рис.2 означает: а) в сегментации $I_{w_{j+p+1}}$ цепочка $v = a_{j+m} \dots a_{j+m+r}$ либо является частью информативной зоны, либо входит в состав исходной цепочки для $I_{w_{i+p+1}}$; б) в сегментации $I_{w_{j+p}}$ цепочка — часть неинформативной зоны; в) сегментации $I_{w_j}, \dots, I_{w_{j+p}}$ совпадают и отличны от сегментации $I_{w_{j-1}}$ ($1 < j < k$). Иначе говоря, структура, приведенная на рис.2, позволяет определить, какие элементы цепочки \tilde{w} однозначно предсказывают появление цепочки v .

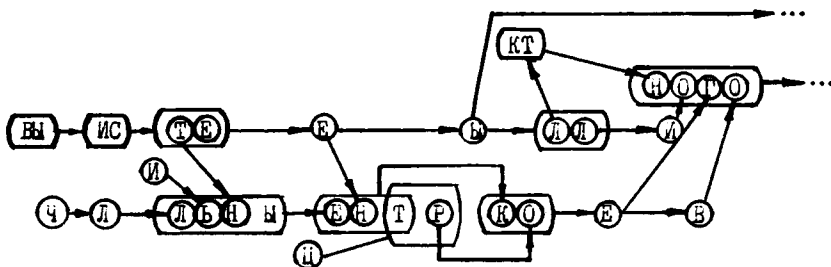


Рис. 2

Если из текста не были выброшены пробелы (как в приведенных примерах), то можно использовать их при интерпретации результата.

Это позволит: а) выделять связи между морфами одного слова и морфами соседних слов; б) проследить согласованность соседних слов в разных формах, выделяя тем самым устойчивые для данного текста конструкции слов.

Сложность применения описанного в статье варианта метода реконструкции к анализу больших текстов ограничивает возможности его использования при исследовании языковых особенностей текстов того или иного подязыка естественного языка (основное назначение метода [1]). Метод, описанный в [1], требовал анализа больших текстов и его применение было затруднено сложностью введения этих текстов в ЭВМ. Модификация, предлагаемая в данной статье, дает наглядные результаты на текстах небольшой длины (3-5 тысяч символов или 1-2 страницы), что дает возможность ее более широкого использования при исследовании структуры отдельных участков текста.

Л и т е р а т у р а

1. ТИМОФЕЕВА М.К. О методике индуктивной реконструкции грамматики. -Новосибирск, 1983. Б.и. - 18 с. (Препринт/ИМ СО АН СССР: № 40)

2. ГУСЕВ В.Д., КУЛИЧКОВ В.А., ТИТКОВА Т.Н. Анализ генетических текстов. I. 1-граммные характеристики. -В кн.: Эмпирическое предсказание и распознавание образов (Вычислительные системы, вып. 63). Новосибирск, 1980, с. 11-33.

3. ГУСЕВ В.Д., КОСАРЕВ Ю.Г., ТИТКОВА Т.Н. О задаче поиска повторяющихся отрезков текста. -В кн.: Вычислительные системы, вып. 62. Ассоциативное кодирование. -Новосибирск, 1975, с.49-71.

4. ТИМОФЕЕВА М.К. Применение R-технологии программирования для организации больших словарей в памяти ЭВМ. -В кн.: Автоматизированные системы управления ВУЗом. Новосибирск, 1978, с.57-66.

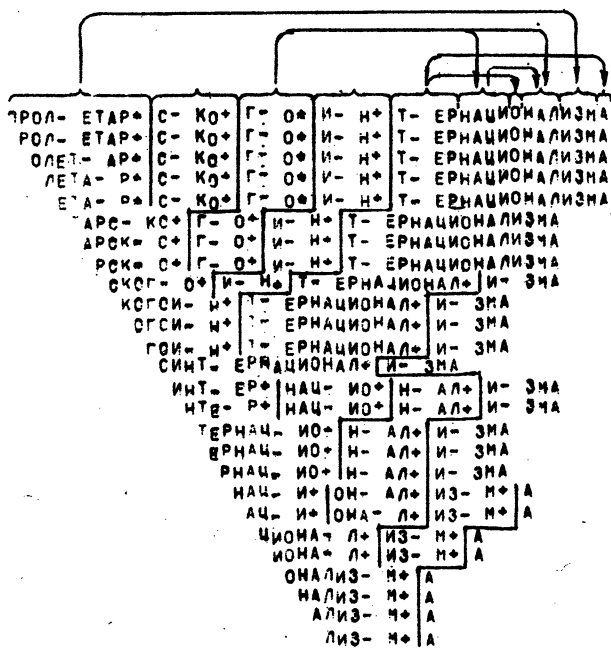
Поступила в ред.-изд.отд.
21 августа 1985 года

Множество сегментаций, полученное при анализе цепочки W
из текста научно-технических рефератов справа налево

ЯИНАВ- 0+	3- ЫЛОП+	0- ГОНВИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ
ИНАВ- 0+	3- ЫЛОП+	0- ГОНВИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ
НАВ- 0+	3- ЫЛОП+	0- ГОНВИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ
АВОЗ- ЫЛОП+	0- ГОНВИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ	
ВОЗЬ- ЛОР+	0- ГОНВИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ	
ОЗЬ- ЛОР+	0- ГОНВИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ	
ЗЬ- ЛОР+	0- ГОНВИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ	
ЫЛОГО-	ГОНВИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ	
ЛОГО-	ГОНВИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ	
ОЛО-	ГОНВИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ	
ЛОГ- ОН+	В- ИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ	
ОГОНВ-	ИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ	
ГОНВ-	ИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ	
ОНВИ-	ИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ	
НВИ-	ИТКЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ	
ВИТ-	КЕЛ+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ	
ИТКЕ-	Л- ОК+	Л- ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ	
ИТКЕЛ-	ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ		
КЕЛЛ-	ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ		
ЕЛЛ-	ОК+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ		
ЛОЛ-	К+	Ы- РТНЕЦЕШЬЛЕТИЛСИЧЬВ		
ЛОКМ-	РТНЕЦЕШЬЛЕТИЛСИЧЬВ			
ОКМ-	РТНЕЦЕШЬЛЕТИЛСИЧЬВ			
КМР-	ТНЕЦЕШЬЛЕТИЛСИЧЬВ			
МРТ-	НЕЦЕШЬЛЕТИЛСИЧЬВ			
РТН-	ЕЦ+ Е- ШЬЛЕТИЛСИЧЬВ			
ТНЕЦЕ-	ШЬЛЕТИЛСИЧЬВ			
НЕЦЕ-	ШЬЛЕТИЛСИЧЬВ			
ЕЦЕ-	ШЬЛЕТИЛСИЧЬВ			
ЦЕШ-	ШЬЛЕТИЛСИЧЬВ			
ЕШЬ-	ЛЕТИЛСИЧЬВ			
ШЬ-	Л+ Е- ТИЛСИЧЬВ			
ШЬ-	Л+ Е- Т+ И- ЛСИЧЬВ			
ЬЛЕТИ-	ЛСИЧЬВ			
ЛЕТИ-	ЛСИЧЬВ			
ЕТИЛ-	СИЧЬВ			
ТИЛ-	ИЧЬВ			
ИЛ-	ИЧЬВ			
ЛСИ-	Ч+ И- В			
	СИЧЬ- В			
	ИЧЬ- В			

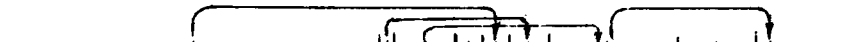
RF1421 = STEP WAS EXECUTED - COND CODE 4000
 123731 STEP /GO / START 85248.2341
 513741 STEP /GO / STOP 85248.2342 CPU (MIN 20.70SEC MAIN 94K

Множество сегментаций, полученное при анализе цепочки 8
из общественно-политического текста слева направо



EF1421 STEP WAS EXECUTED - COND CODE 4850
 EF3771 STEP /GO / START 85246.2335
 EF13741 STEP /GO / STOP 85246.2340 CPU 1MIN 25.38SEC MAIN 82K

Множество сегментаций, полученное при анализе цепочки
из генетического текста SV4 ϕ слева направо



TGGTTG- CTGACSTAATTGAGATGCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 GGTGCT- GACSTAATTGAGATGCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 CTTGCTG- ACTAATTGAGATGCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 TTGCTGA- STAATTGAGATGCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 TGCTGA- STAATTGAGATGCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 GCTGACT- AATTGAGATGCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 CTGACT- AATTGAGATGCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 TGAATA- TTGAGATGCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 GACTAA- TTGAGATGCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 ACTAAT- TGAGATGCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 CTAATTGA- GATGCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 TAAATGA- GATGCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 AATTGAG- ATGCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 ATTGAG- ATGCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 TTGAGAT- GCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 TGAGAT- GCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 GAGAT- GCATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 AGATGC- ATGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 GATGCA- TGCTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 ATGCATG- CTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 TGCATG- CTTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 GCATGC- TTTGCATACTTCTGCGCTGCTGGGGAGCCTGG
 CATGCTTT- GCATACTTCTGCGCTGCTGGGGAGCCTGG
 ATGCTTTG- CATACTTCTGCGCTGCTGGGGAGCCTGG
 TGCTTTG- CATACTTCTGCGCTGCTGGGGAGCCTGG
 GCTTTG- CATACTTCTGCGCTGCTGGGGAGCCTGG
 CTTTGCAT- ACTTCTGCGCTGCTGGGGAGCCTGG
 TTTGCAT- A+|C- TTTGCGCTGCTGGGGAGCCTGG
 TTGCATAC- T+|T- CTGCGCTGCTGGGGAGCCTGG
 TGCATAC- T+|T- CTGCGCTGCTGGGGAGCCTGG
 GCATAC- T+|T- CTGCGCTGCTGGGGAGCCTGG
 CATACTT- CTGCGCTGCTGGGGAGCCTGG
 ATACTTC- TGCGCTGCTGGGGAGCCTGG
 TACTTC- T+|GCC- TGCTGGGGAGCCTGG
 ACTTCTGCC- TGCTGGGGAGCCTGG
 CTTCTGCC- TGCTGGGGAGCCTGG
 TTCTGCC- TGCTGGGGAGCCTGG
 TCTGCC- TGCTGGGGAGCCTGG
 CTGCGTG- CTGGGGAGCCTGG
 TGCGTGC- TGGGGAGCCTGG
 GCCTGCT- GGGGAGCCTGG
 CCTGCTGG- GGAGCCTGG
 CTGCTGGG- GAGCCTGG
 TGCTGGGG- AGCCTGG
 GCTGGGG- AGCCTGG
 CTGGGGAG- CCTGG
 TGGGGAG- CCTGG
 GGGGAGC- CTGG
 GGGAGC- CTGG
 GGAGCC- TGG
 GAGCC- T+|GG

IEF1421 - STEP NAS EXECUTED - COND CODE 5200
 IEF3731 STEP /GO / START 85240.2347
 IEF3741 STEP /GO / STOP 85240.2349 CPU 1MIN 23.00SEC MAIN 94K