Алгоритмические аспекты анализа символьных последовательностей

Мирошниченко Любовь Александровна
Институт математики СО РАН
luba@math.nsc.ru

Объект исследования: символьные последовательности.

Σ – непустое конечное множество символов (алфавит);

$T = t_1 t_2 \ldots t_i \ldots t_N \quad (t_i \in \Sigma,\ 1 \le i \le N)$ – последовательность символов, цепочка символов, текст, строка, слово.

Примеры:

- двоичные последовательности ($|\Sigma| = 2$);
- последовательность действий, траектория движения;
- тексты программ;
- тексты естественного языка;
- музыкальные тексты (песенные мелодии);
- древнерусские церковные песнопения;
- ДНК, РНК ($|\Sigma| = 4$);
- аминокислотные послед. ($|\Sigma| = 20$);
- порядки генов;
- порядки дисков политенных хромосом …

Основные задачи анализа символьных последовательностей:

- поиск образцов;
- восстановление структуры текста: выявление повторов (периодичностей, симметрий …);
- сложность текста

Сравнение последовательностей

Анализ параллельных текстов

Поиск образцов

$P = p_1 p_2 \ldots p_m$ – образец, $T = t_1\, t_2 \ldots t_N$ – текст, $t_i \in \Sigma$, $p_j \in \Sigma$, $1 \le i \le N$, $1 \le j \le m$, $m < N$

Задача: обнаружить все вхождения образца *P* в текст *T*.

Обобщения:

- Образцы с джокерами: $p_j \in \Sigma \cup \{x\}$, x – любой символ
- Образцы, позиции которых заданы множествами символов:
  P = a- [ag]-c-[cg]-[acg]-[ct]-a, P = a-[ag]-c-[cg]-¬t-*x*-a
- Поиск образца с «ошибками»
- Поиск множества образцов (попевки)
- Поиск множества образцов, позиции которых заданы множествами символов. Кол-во образцов ~ 10^5. Алгоритмы используют «ядерную» стратегию в сочетании с построением ДКА.

  Бачинский А.Г., Гусев В.Д., Немытикова и др. Новая версия банка образов PROF–PAT 1.0: технология формирования и программа быстрого поиска образов в аминокислотных последовательностях // Молекулярная биология, 1996, Т. 30, вып. 6, С. 1409–1417
- Образцы с переменными
  P = ab*XX*c*X* : abttct; ababbabbcabb
- Параметризованные образцы: 2 алфавита: Σ и Π:
  Образцы abc*X*bb*YY*cc*Z* и abc*Z*bb*XX*cc*Y* π-согласованы

Повтор – пара совпадающих фрагментов текста

Классификация повторов

По характеру расположения в тексте:

- разнесенные … agttc … agttc…
- тандемные … agttcagttc…
- с наложением : … agttcagttcagttc …

По наличию искажений:

- совершенные … agttc … agttc …
- несовершенные … agttc … aattc … , … agttc … agtttc… ,
- с точностью до агрегирования: … agttc … gatct …({a,g}→u, {c,t}→y).
- с точностью до подстановки на элементах алфавита:
  … agttc… tcaag… (a ↔ t, c ↔ g); секвентные переносы в музыке;

По направлению считывания

- прямые : … agttc … agttc …
- симметричные: … agttc … cttga…
  инвертированные: … $\overrightarrow{\text{agttc}}$… $\overleftarrow{\text{gaact}}$ …

Представление текста в терминах повторов

Σ – конечный алфавит;

$S = s_1 s_2 s_3 s_4 s_5 \ldots s_N$ – текст, составленный из элементов Σ;

$S[i:j] = s_i \ldots s_j$ – фрагмент текста ($1 \le i < j \le N$).

***l*-грамма** – фрагмент текста длины l ($S[i : i + l - 1]$).

$$S = s_1 s_2 s_3 s_4 s_5 \ldots s_N$$

Полное число l-грамм: $N - l + 1$.

Число различных l-грамм: $M_l \le N - l + 1$.

$l_{max}(S)$ – наибольшее l, при котором в S есть ***повторяющиеся* *l*-граммы**.

Частотная характеристика l-го порядка текста S –
множество элементов $\Phi_l(S) = \{\phi_{l1}, \phi_{l2}, \ldots, \phi_{l M_l}\}$
где $\phi_{l i}$ $(1 \le i \le M_l)$: « i-я l-грамма – x_i, ее частота в тексте – $F_l(x_i)$ ».

Совокупность частотных характеристик

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

называется ***полным частотным спектром*** *текста S*.

Наиболее важные параметры частотного спектра:

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

$\boldsymbol{M_l}$ — размер «словаря» l-грамм.

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

E_l^k — Число различных l-грамм, встречающихся в тексте ровно k раз

E_l^0 — число l-грамм, ни разу не встретившихся в тексте.

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

Позиционные аномалии:

- кластеры l-грамм
- сверхравномерное распределение
- «изолированные» точки
- «гэпы»

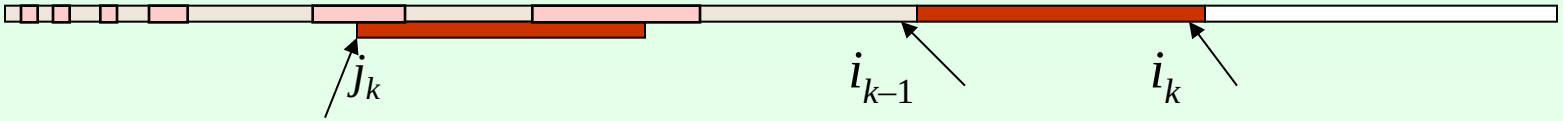
Сложность по Лемпелю и Зиву

Сложность $c_{LZ}(S) = m$ последовательности S: число шагов процесса ее «порождения». Каждый шаг: копирование максимально длинного фрагмента из синтезированного префикса и/или генерация символа.

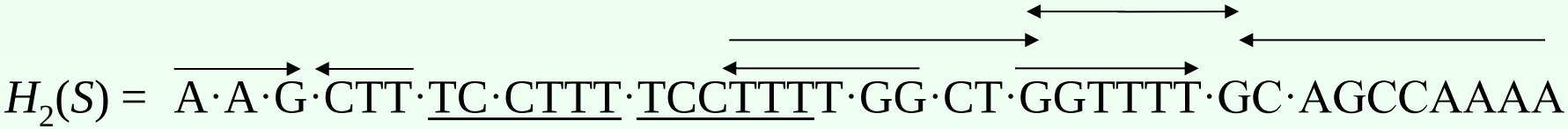
Сложностное разложение S есть конкатенация

$$H(S) = S[1:i_1]\,S[i_1+1:i_2]\ldots S[i_{k-1}+1:i_k]\ldots S[i_{m-1}+1:N],$$

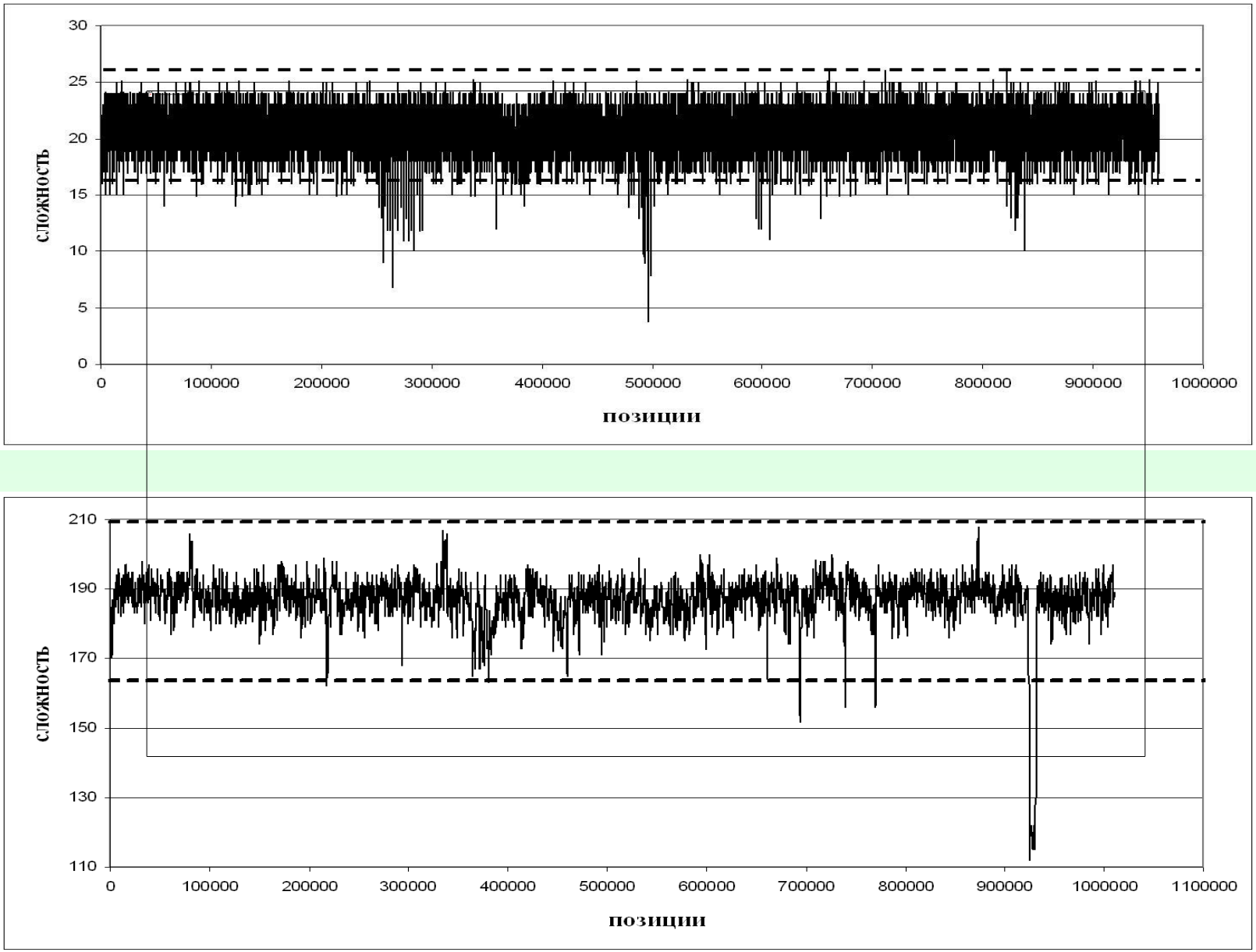
где $S[i_{k-1}+1 : i_k] = S[j_k : j_k + l_k - 1]$ — фрагмент, синтезируемый на k-м шаге, $j_k < i_{k-1}+1$



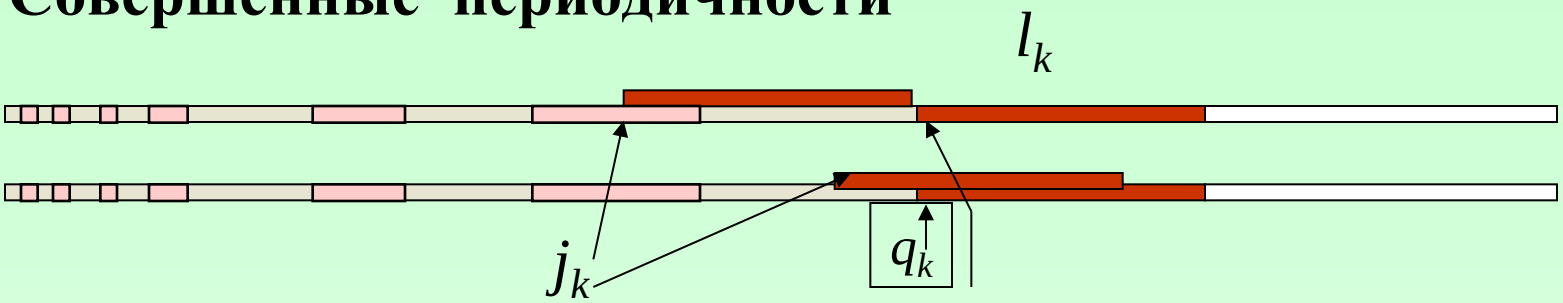
ДНК-ориентированный вариант меры сложности наряду с прямыми повторами учитывает также симметричные и комплементарные (прямые и симметричные).



Профили сложности генома микоплазмы «R» при W = 60 и W = 1000



Совершенные периодичности



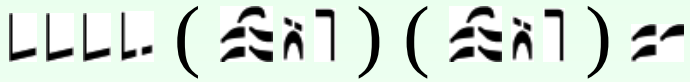
Если $j_k + l_k \geq q_k$, имеет место периодичность.

Длина периода $p \leq q_k - j_k$; кратность $\geq \lfloor l_k / p \rfloor + 1$.

Комбинации периодичностей (компаунды):

(acacacacacacacacac)(atatatatatatat)

(agagagagag)(agatagatagatagatagat)



серия стопиц, стопица с очком (змийца со статьей, голубчик борзый малый, палка)^2, статья мрачная

- В темном лесе, в темном лесе,
- В темном лесе, в темном лесе,
- За лесью, за лесью,
- Распашуль я распашуль я
- Распашуль я распашуль я
- Пашенку, пашенку.
- Я посею, я посею,
- Я посею, я посею лен,
- Конопель, лен зеленой.

Регулярные повторы

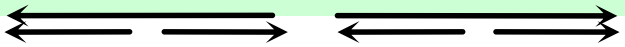
gttttggggttgtacaattattttgttaagtaaaac aaatgataaataacgcttaacttgcttact
gttttggggttgtacaattattttgttaagtaaaac cctataaacaaatcaggattatatgtacta
gttttggggttgtacaattattttgttaagtaaaac ttaagtcaagatttttaataccagggtgca
gttttggggttgtacaattattttgttaagtaaaac tccatatttttccttactattactatgct
gttttggggttgtacaattattttgttaagtaaaac acgatttaaaaattatgatataatataac
gttttggggttgtacaattattttgttaagtaaaac tagaatctcttttaaattcccaccaagct
gttttggggttgtacaattattttgttaagtaaaac ctttttttaaattttattattataactttgt
gttttggggttgtacaattattttgttaagtaaaac aattgcatcattaacgttaaagacgttact
gttttggggttgtacaattattttgttaagtaaaac aatcaaacaactcgctttctaaatcatcaa
gttttggggttgtacaattattttgttaagtaaaac tttgagcataatggcgctttgagtctttag
gttttggggttgtacaattattttgttaagtaaaac aatgcagatttaaaagattcaggaacgatt
gttttggggttgtacaattattttgttaagtaaaac attagccccacaattatattaacctccct

геном "Mycoplasma synoviae 53" (ID AE017245), поз. 690229

CRISPRs (*Clustered Regularly Interspaced Short Palindromic Repeats*)

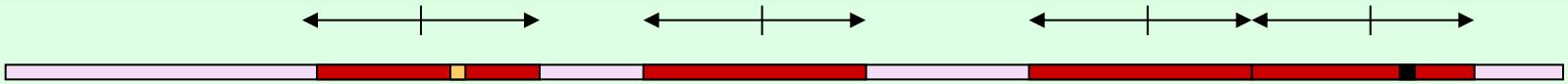
Тексты песен: куплет – припев…

Фрактальные структуры:



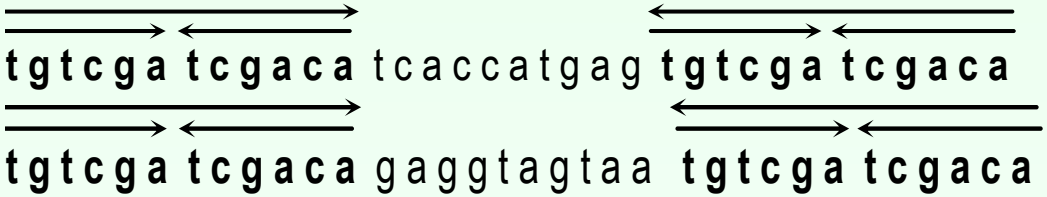
Класс 1 : (ag-ga)(ag-ga)(ag-ga), поз. 5870, 25/80 (низковир.);
(gtg)(gtg)(gtg)(gtg) = V^4, поз. 805, 10/80 (высоковир.);
Класс 2 : ggaagg ggaagg gg = EGEG, поз. 9411, 27/46 + 1/28

Фракталоподобные структуры:



Все ВКЭ: gagctcaaactggagagctc, 146/161
Класс 1 : gttggttgctggttgttg, поз. 308, 70/80
Класс 2 : cggcaccaaccggctcggcggc, поз. 8471, 41/46

Arabidopsis th (4-я хромосома), 194 вхождения:



Комбинированные структуры в геноме ВКЭ

Комбинированными назовем структуры, состоящие из двух разнотипных повторов (прямой + симметричный, прямой + комплементарный симметричный и т.п.).

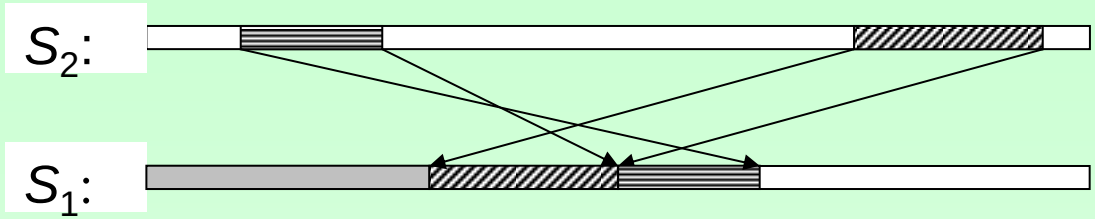
```
Класс 1 (66/80).  Позиции: Xs - 1916; X - 1927; Y - 1944; Y - 1947
       gataacacccaaccccacaatcgaaaccaatggtggtggttt
     Xs = taacaccc   cccacaat = X
      Y =                                tggtggt
      Y =                                   tggtggt
638: L  I  T  P  N  P  T  I  E  T  N  G  G  G
```

Класс 3 (18/28) Позиции: Xs - 752; X - 788; Y_{11} - 766; Y_{12} – 785, Y_{21} - 767; Y_{22} - 807

```
    ggcagtggtagccgttgtgtggttgaccgtggaaagtgtggtgacccgagtcgccgttgtggttgtg
Xs  = cagtggt                                tggtgac = X
Y11 =               gtgtggt
Y12 =                                  gtgtggt
Y21 =                tgtggttg
Y22 =                                                    tgtggttg
250: A  V  V  A  V  V  W  L  T  V  E  S  V  V  T  R  V  A  V  V  V  V
```

```
Класс 4 (7/7)  Позиции : Xs - 6954;  X - 6998; Y1 - 6970; Y2 - 6981
   cagctccagaccaagatccaacagcttgtcaacagcgccgtcgcgactggagcac (51);
Xs = gctccag                                          ctggagc = X
 Y =                  caacagc
 Y =                              caacagc
```

Сравнение символьных последовательностей



- **Сложностное разложение** S_1 по S_2 : покрытие S_1 фрагментами S_2.
- На каждом шаге копируется максимальный фрагмент S_2, совпадающий с префиксом непокрытого участка S_1
- **Сложность c(S_1 / S_2) :** число компонентов в разложении S_1 по S_2

Инверсионное расстояние $d_I(\pi,\sigma)$ между последовательностями π и σ определяется минимальным числом инверсий, переводящих одну из них в другую

$$\begin{pmatrix} 1\,2\ldots i-1 & \underset{\longleftarrow}{i} & i+1\ldots j-1 & \underset{\longrightarrow}{j} & j+1\ldots N \\ 1\,2\ldots i-1 & j & j-1\ldots i+1 & i & j+1\ldots N \end{pmatrix}$$

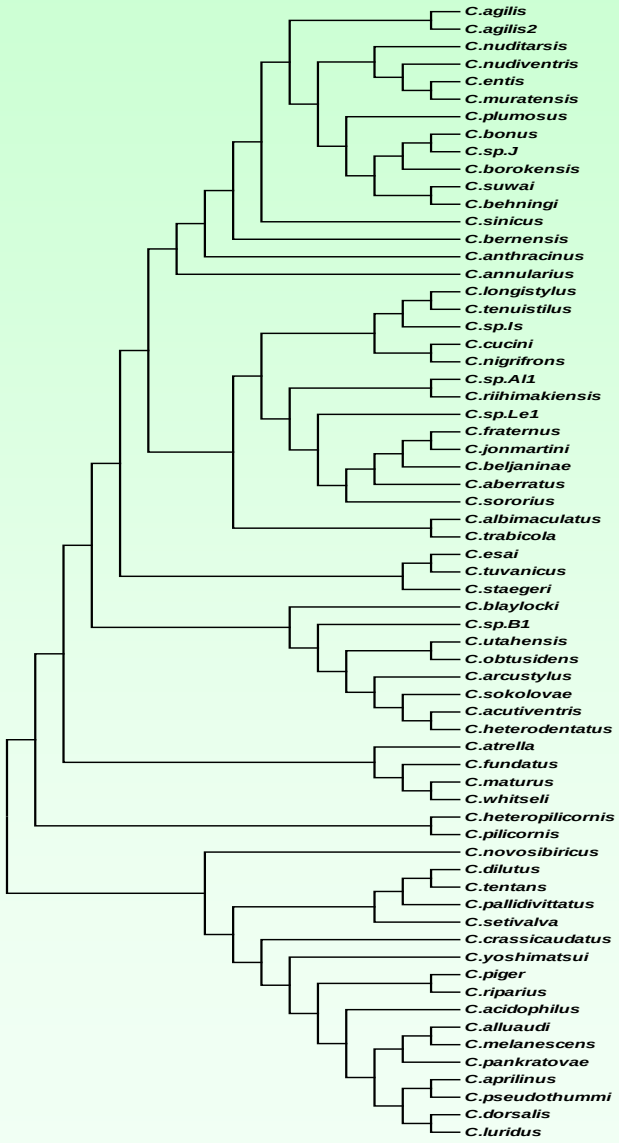
Точки разрыва для перестановок π и σ:

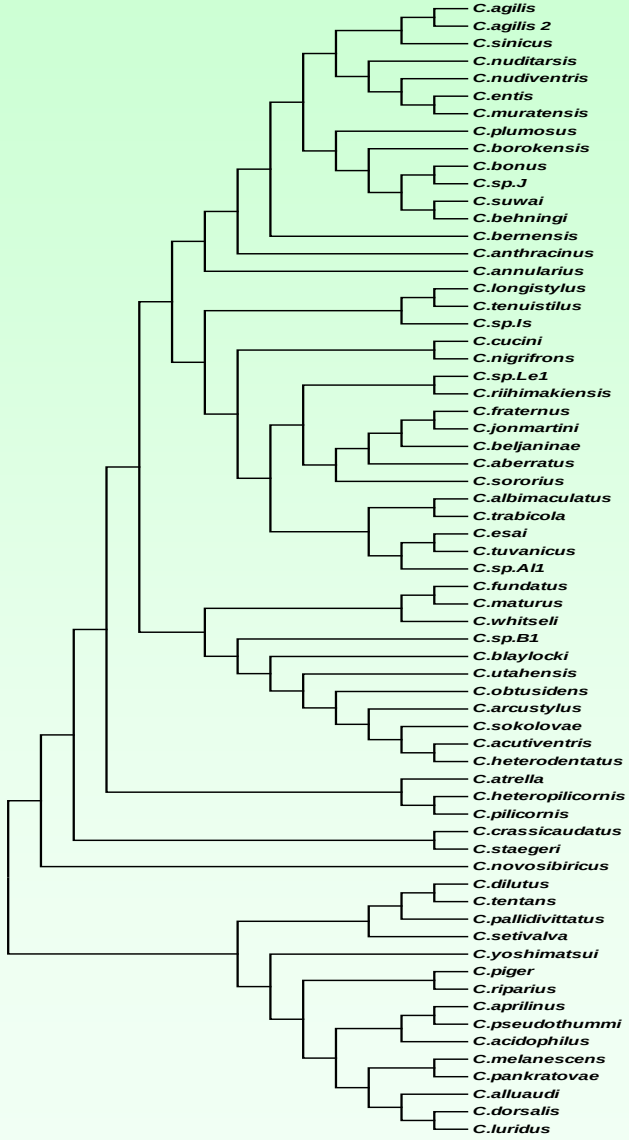
$\pi_i = a$, $\pi_{i+1} = b$, в σ нет биграмм ab и ba

π = 0 | 6 4 | 1 8 5 | 3 | 2 9 7 | 10 $\qquad$ $r(\pi, \sigma) = 5$

σ = 0 | 5 8 1 | 2 9 7 | 6 4 | 3 | 10

Phylogenetic trees for 65 species of the genus *Chironomus*





Сравнение «близких» последовательностей

$K = \{П_1, П_2, \ldots П_k\}$ – множество текстов, разбитое на k классов

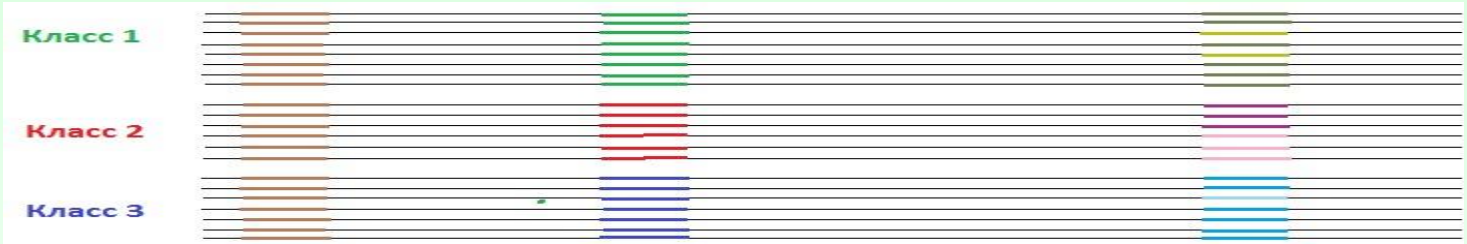
$П_i = \{T_{i1}, T_{i2}, \ldots T_{imi}\}$ $(1 \le i \le k)$, m_i – количество текстов в i-м классе.

$Ф_L(П_i / K)$ – множество **«контрастных» *L*-грамм**, присутствующих в каждом тексте i-го класса и отсутствующих во всех остальных текстах.

Класс 1 ($m_1 = 80$) : $L_{max}(Ф(П_1 / K)) = 17$; $L_{min}(Ф(П_1 / K)) = 6$

Класс 2 ($m_2 = 46$) : $L_{max}(Ф(П_2 / K)) = 35$; $L_{min}(Ф(П_2 / K)) = 7$

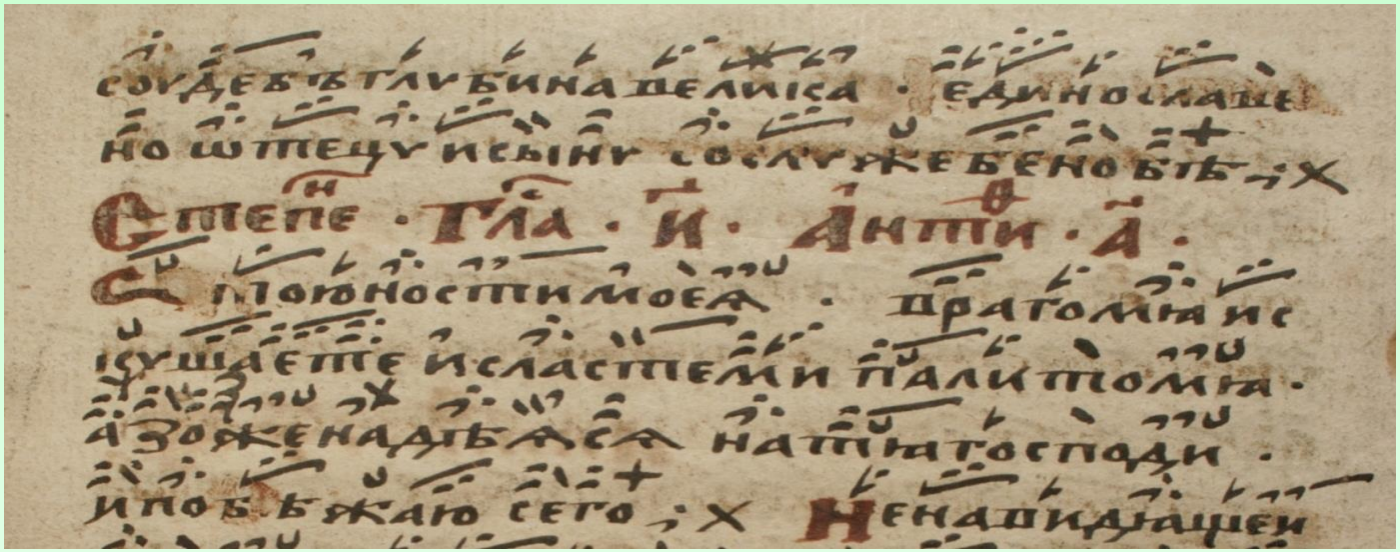
- ***L*-граммно-позиционный «срез»**



Позиция *pos* «классифицирующая», если множества *L*-грамм в текстах разных классов не пересекаются

- **Периодичности**
- **Фрактальные структуры**

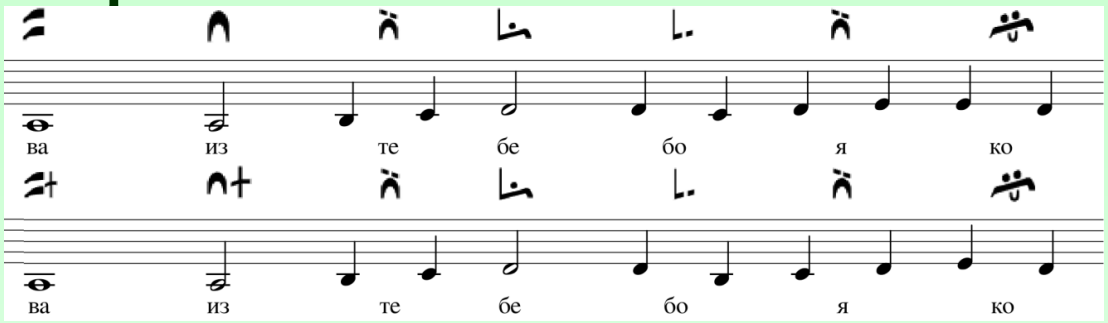
Анализ параллельных текстов



Варианты распева крюка светлого в разных гласах

	473	525	442	506	443	334	321	612
R1= 	408	508	422	484	408	319	298	600
			c - 1			c - 2	c - 1	c - 2
	d - 1			d - 4		d - 11	d - 2	
	e - 26	e - 3		e - 6		e - 28	e - 71	e - 2
	f - 70	f -103	f - 33	f -100	f - 52	f -177	f - 34	f -189
	g -211	g -191	g -112	g -317	g -233	g - 90	g -171	g -260
	a -100	a -206	a -234	a - 57	a -116	a - 11	a - 19	a -142
		b - 5	b - 17		b - 3			b - 5
			C - 25		C - 2			
					D - 2			
R2= 	53	2			16	1		
I1(R2)=2+	g - 53				C - 16			
I2(R2)=1-		f - 2				f - 1		
R3= 	11	14	17	11	13	10	17	8
				d - 1			d - 1	d - 2
						e - 4		
	f - 4	f - 1		f - 3	f - 3		f - 5	f - 1
	g - 7	g - 2	g - 2	g - 7	g - 5		g - 11	g - 1
		a - 11	a - 15		a - 1	a - 6		a - 4
					b - 1			
					C - 3			
R4= 	1		3	9	5	4	4	3
						d - 2		
			f - 1	f - 5	f - 4	f - 1	f - 3	f - 2
	g - 1			g - 4			g - 1	
			a - 2		a - 1	a - 1		a - 1
R5= 		1						
I(R5)=1-		a - 1						
R6= 				g - 2			g - 2	
R7= 					a - 1			
R8= 								1
I(R8)=2-								g - 1

Вариативность. Сравнение одноименных песнопений в разных певческих книгах



знамя	нота	
+	+	1218
+	–	118
–	+	192
–	–	127

> Sol_619_647; glas1; pesn 12 - > Sol_618_644; glas1; pesn 12

```
(r0201-d2по)   (r0211-d4H4ро) (-0511-c4d4жде)(s0121-e2стве)(n0201-d2я)
(r0201-d2по)   (r0211-d4c4рож)(-0511-c4d4де) (s0111-e2стве)(r0201-e2я)
     +++            +--            ++-            -++            +-+
(r0201-d2ко)   (r0201-d2пре)   (r0201-d2жде)   (r0201-d2рож) (g0402-c2де)
(r0201-e2ко)   (r0201-e2пре)   (r0201-e2жде)   (n0201-d2рож) (g0401-c2де)
     +-+            +-+            +-+            +++            -++
(n0111-d2ства)(r0211-d4H4пре)(-0511-c4d4бы) (s0121-e2ла)   (g0302-c4H4е)
(n0201-d2ства)(r0211-d4H4пре)(-0511-c4d4бы) (s0111-e2ла)   (g0302-c4H4е)
     -++            +++            +++            -++            +++
```

73% совпадений для 21 пары одноименных песнопений (глас 1)

Точность дешифровки коррелирована с оценкой вариативности.

Алгоритмы дешифровки знаменных песнопений.

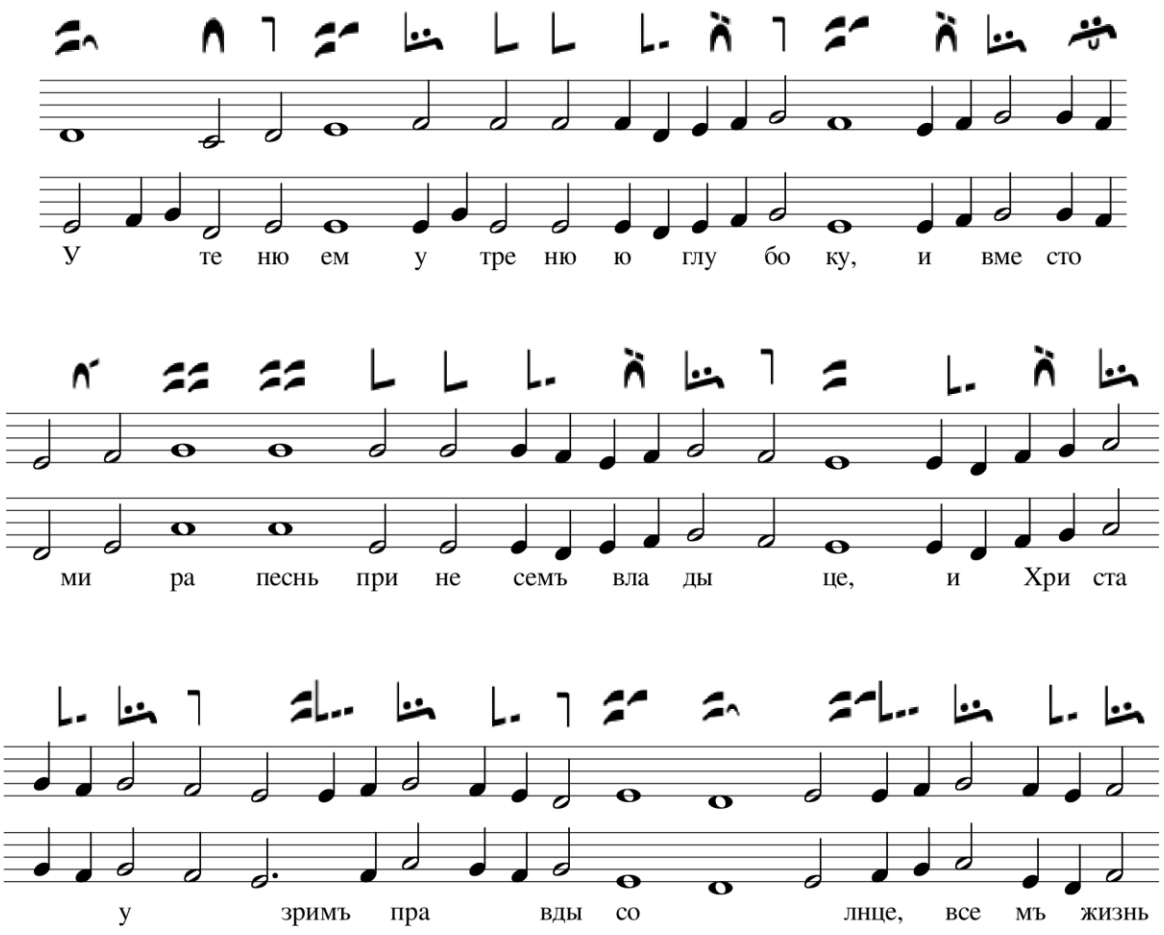
- Нулевое приближение.
- 3-граммный подход (учет минимального контекста)
- Покрытие длинными фрагментами
- Внутригласовые инварианты (ВИ) и квазиинварианты (КВИ)
- Покрытие попевками в сочетании с ВИ/КВИ
- Связные пути многоуровнего графа

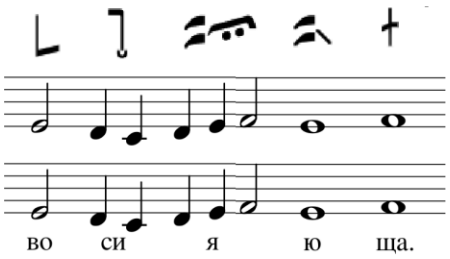
$k = n_+ / N$, где n_+ – число «правильно» интерпретированных знамен,

N – суммарная длина песнопений.

Подход	**Гласы**							
	1	2	3	4	5	6	7	8
1	0.59	0.49	0.5	0.5	0.5	0.42	0.54	0.43
2	0.64	0.58	0.62	0.6	0.58	0.52	0.71	0.48
3	0.64	0.63	0.63	0.63	0.63	0.52	0.78	0.5
4	0.68	0.63	0.65	0.69	0.67	0.56	0.79	0.52
5	0.69	0.64	0.66	0.7	0.67	0.57	0.83	0.51
6	0.68	0.63	0.67	0.57	0.61	0.55	0.76	0.53

Исходный текст и дешифровка





Ирмологий, глас 1, песн. 34.
45 знамен
Точность реконструкции 24/45 = 0.53
3 ошибки в восстановлении ритма

Благодарю за внимание!