СТРУКТУРНЫЕ АНАЛОГИИ В СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ РАЗЛИЧНОЙ ЯЗЫКОВОЙ ПРИРОДЫ

Гусев В.Д., Мирошниченко Л.А., Саломатина Н.В.

Новосибирск, Институт математики СО РАН

Объект исследования

Σ – непустое конечное множество символов (алфавит);

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

Примеры:

- ДНК, РНК (|**Σ**| = 4); аминокислотные послед. (|**Σ**| = 20);
- порядки генов (|**Σ**| ~ 10^4); порядки дисков политенных хромосом;
- музыкальные тексты (песенные мелодии);
- знаменные песнопения;
- тексты программ;
- слова, предложения,…, тексты естественного языка;
- последовательности действий животных;
- формальные последовательности.

Повтор – структурообразующий элемент

Повтор (в широком смысле) — пара фрагментов текста, совпадающих с точностью до переименования элементов алфавита и (или) изменения направления считывания.

Типы повторов:

– разнесенные ... **AGTTC** ... **AGTTC** ...

– тандемные ... **AGTTCAGTTC**...

– симметричные: ... **AGTTC** ... **CTTGA** ...

– прямые комплементарные: ... **AGTTC** ... **TCAAG** ...

– инвертированные: ... **AGTTC** ... **GAACT** ...

– с произвольным переименованием элементов:

```
if j  > 1 then k := k + 1 else h := h*2;
if pr2 > 1 then tam := tam + 1 else tut := tut*2
```

Сложность последовательности

I. определения учитывающие частоту встречаемости символов или относительно коротких *l*-грамм (энтропийные меры, SIMPLE…)

II. разнообразие *l*-граммного состава (комбинаторная сложность, лингвистическая …)

III. определения, основанные на учете длинных повторов (грамматическая сложность, Лемпеле-Зив-подобные меры…).

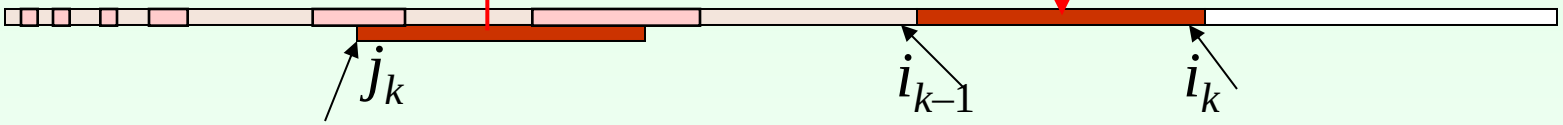
Представление текста в виде конкатенации повторяющихся фрагментов

Сложность последовательности ***S*** по Лемпелю и Зиву измеряется числом шагов порождающего ее процесса. Каждый шаг: копирование максимально длинной "заготовки" из уже синтезированной части текста или генерация символа.

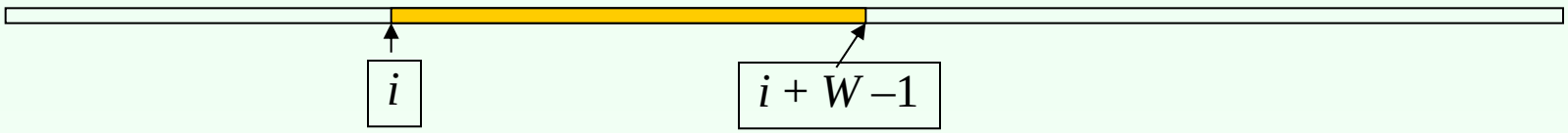
Сложностное разложение *S* есть конкатенация фрагментов

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

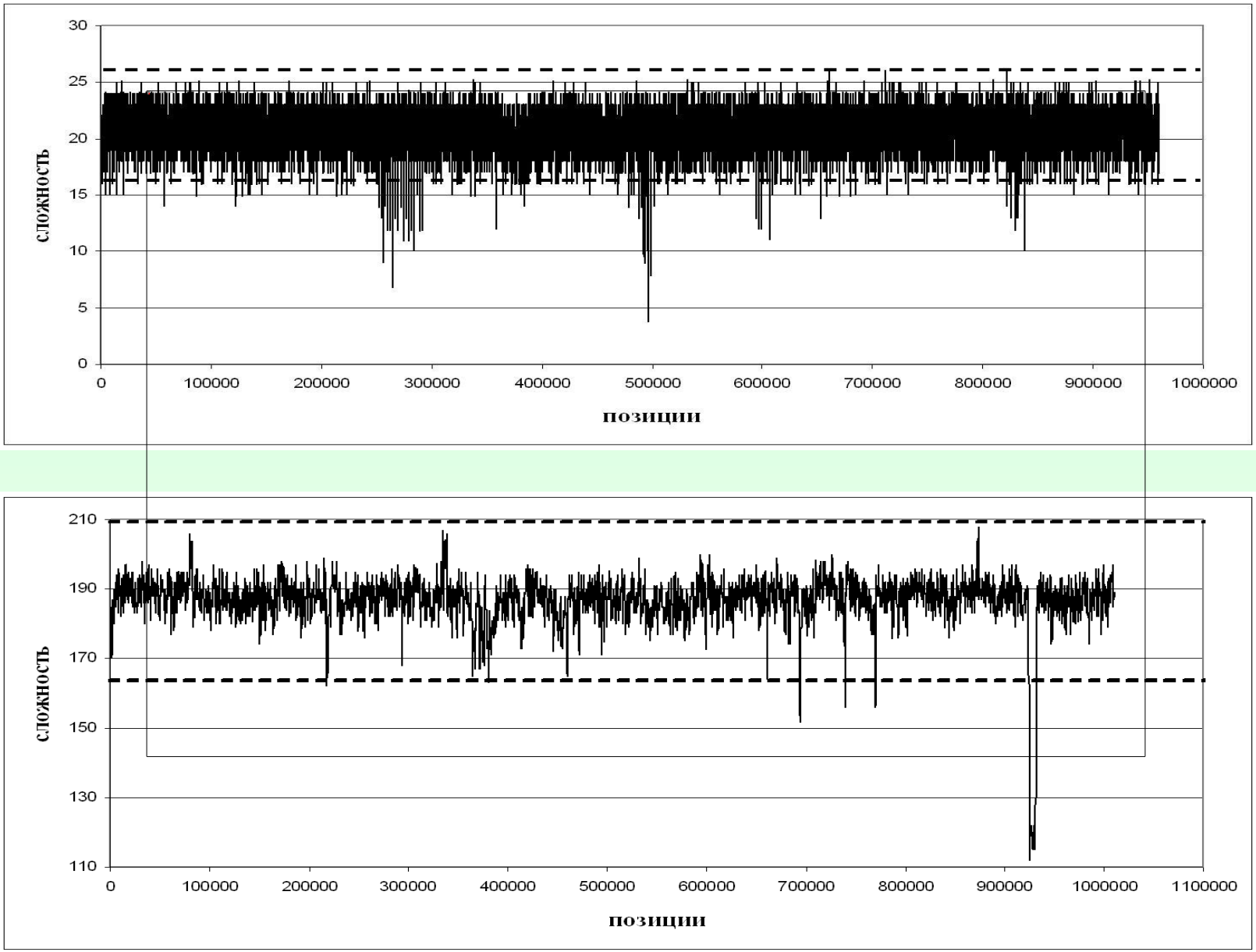
$C_{LZ}(S) = m$ — сложность (число шагов процесса).



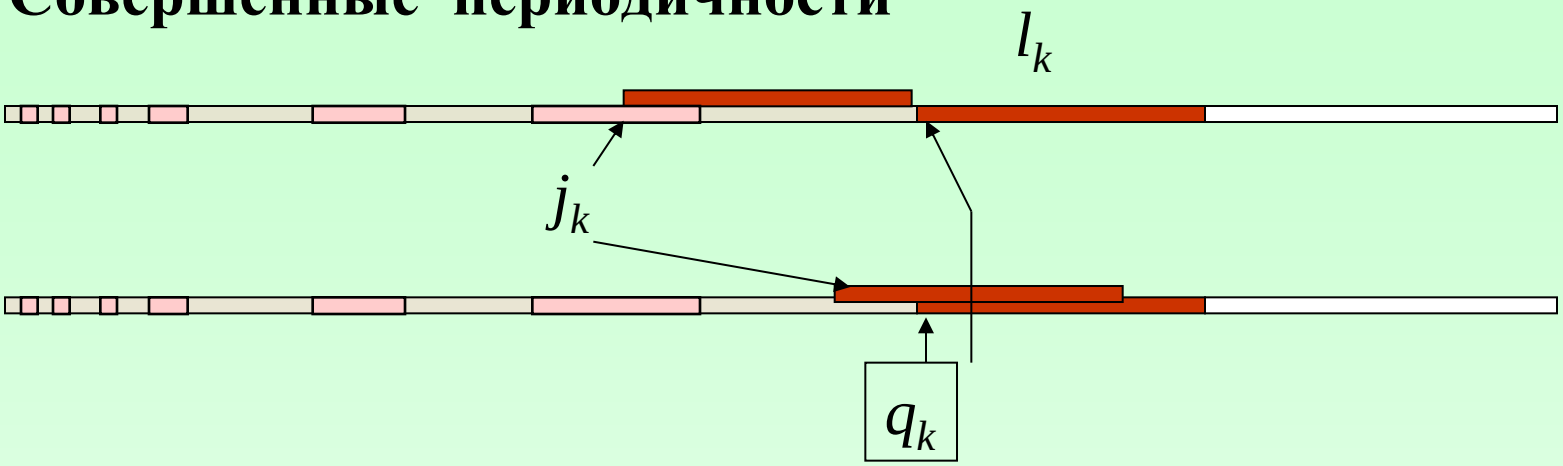
$c_i = C(S[i : i + W - 1])$



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

$H(S) =$ c · a· a · c · ca· t · g · at · $(\text{ccatgat})^4$ · at;

l_k	1	1	1	1	2	1	1	2	28	2	$c(S)=10$
j_k	0	0	2	1	1	0	0	6	4	37	

Совершенные периодичности

- ДНК-последовательности (p от 1 до 10^4)
- Сказки, тексты песен.

Дождливым вечером,
Вечером, вечером …
Пора в путь-дорогу,
Дорогу дальнюю, дальнюю,
Дальнюю идём.,

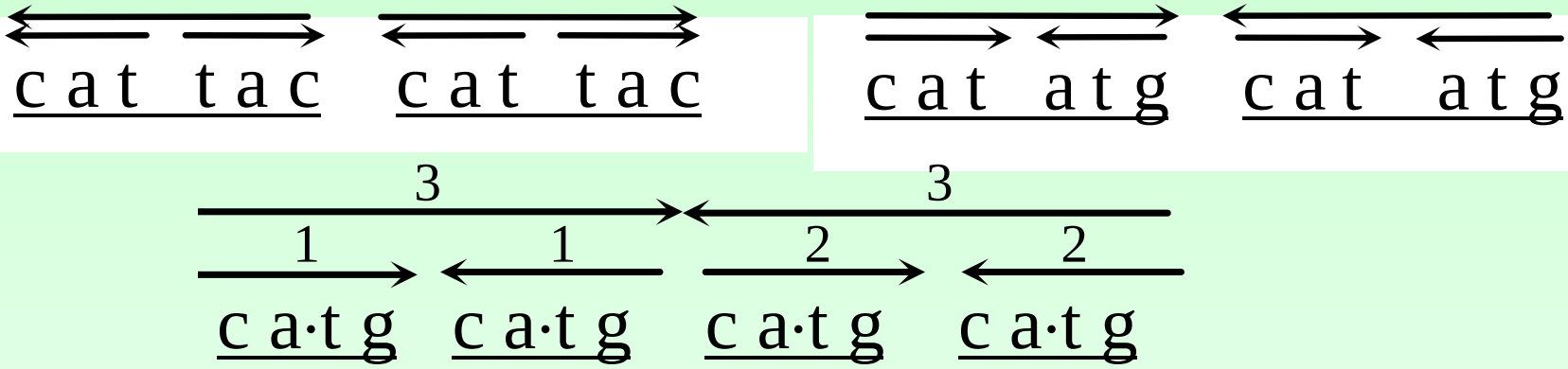
В чаще чаще меньше пищи, значит
в чаще чаще чище. (Б. Заходер)

- Знаменные песнопения

()^2 ()^2 ()^2

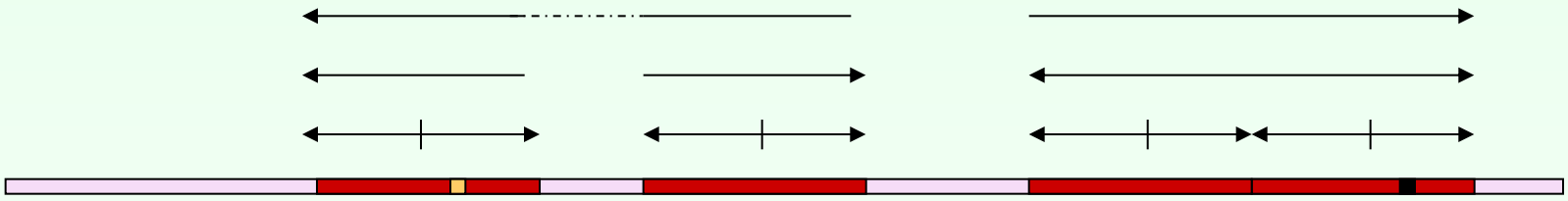
Локальные фракталы в ДНК-последовательностях:

фрагменты текста, обладающие свойством самоподобия.



Фракталоподобные структуры – структуры, в которых позиционно сближены палиндромы

agagaagactagattcaagatcaga



Комбинации периодичностей (компаунды)

(acacacacacacacacacacacacacacacac)(atatatatatatatatatatatatatat)

(agagagagagagag)(agatagatagatagatagatagatagatagatagatagat),

(tctctctctctctctctctc)(tatatatatatatatatatatata)(cacacacacacaca)

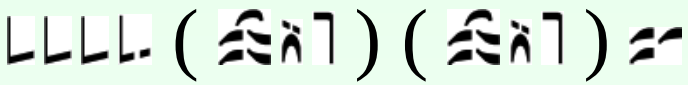
(catgcatgcatgcatg[catgca)acatgcaaca]

Где грань между компаундами и несовершенными периодичностями?

(atatac-atatac-atata)tatatatgtg[tatatatatacata-tatatatatacata-tatata] или $(at)^{30}$?

Комбинации периодичностей (компаунды)

- 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.



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

Периодичности со сложной структурой

MLTR (Multiple Length Tandem Repeat):

$(Xx^n)^m$, где X и x – фрагменты текста, $X \neq x$, $n > 1$, $m > 1$

((gacctttgg) – ac-ac-ac-ac)((gacctttgg) – ac-ac-ac-ac)

VLTR (Variable Length Tandem Repeat) или Nested Tandem Repeat
$x^{n1}Xx^{n2}X\ldots Xx^{nl}$, где $l > 2$, $ni \geq 1$ для $i = 1,\ldots,l-1$ и по крайней мере одно $ni \geq 2$

MPTR – Multi-Periodic Tandem Repeat [Hauth]

	x_1	x_2	x_1	x_3	
S =	(cagta)	(cagca)	(cagta)	(caaca)	$d(S,(x_1)^{12}) = 9$
	(cagta)	(cagca)	(cagta)	(caaca)	$d(S,(x_1x_2)^6) = 3$
	(cagta)	(cagca)	(cagta)	(caaca)	$d(S,(x_1x_2\,x_1x_3)^3 = 0$

- Hauth, A.M. and Joseph, D.A. *Beyond Tandem Repeats: Complex Pattern Structures and Distant Regions of Similarity.* Bioinformatics. 2002 Jul;18 Suppl 1:S31-7.
- Matroud A.A, Hendy M. D. , Tuffley C. P. *NTRFinder: a software tool to find nested tandem repeats,* Nucleic Acids Res., Feb 2012; 40(3): e17

Кумулятивные структуры

A + BA + CBA…или A + AB + ABC… или … ABC + AB + A

A =	Вот дом, который построил Джек.
BA=	А это пшеница, которая в тёмном чулане хранится в доме, который построил Джек.
CBA=	А это весёлая птица – синица, которая часто ворует пшеницу, которая в тёмном чулане хранится в доме, который построил Джек

Репка, колобок …

Вариационность типа прорастания (музыкальные тексты)

Иерархическая дупликация (ДНК):

```
21479 g a t t t g - - g g g a c g a a a a c c a c a t c g t c g t t t
      | | | |         | | | | | | | | | | |   | | |     |     | | | |
21509 g a t t a c - - g g g a c g a a a a c a a c a - g g c a g t t t c
      | | | | | |     | |   | |   | | |   |   | |
21539 g a t t a c - - g g c a c c a a a t c g a c g
        | |   | |     | |
21560 a a t a a c a c g g g g
```

Регулярные повторы

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

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

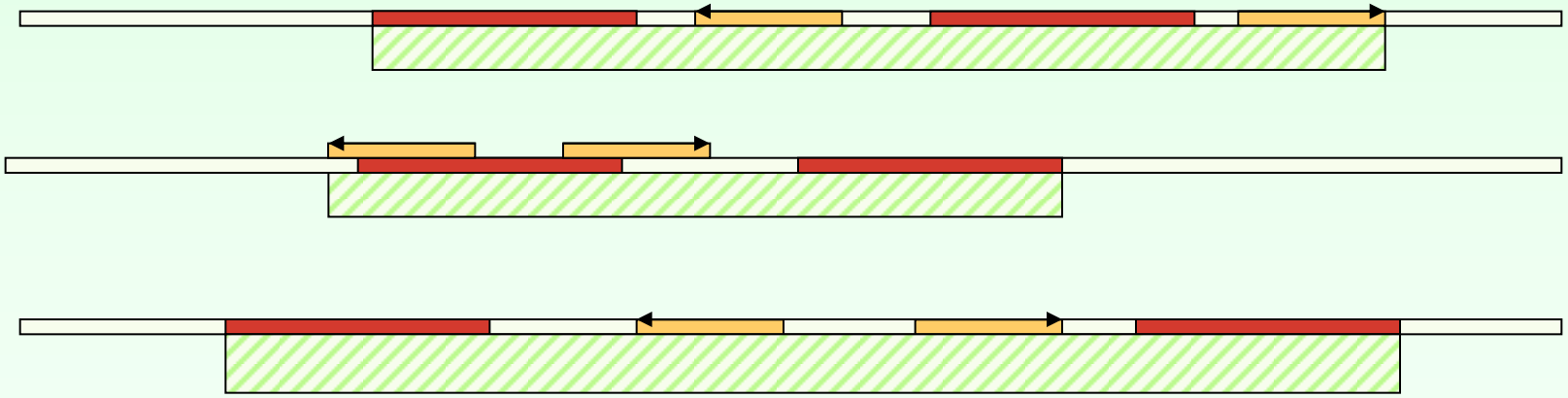
CRISPRs (*Clustered Regularly Interspaced Short Palindromic Repeats*)

Тексты песен: куплет – припев…

Комбинированные структуры

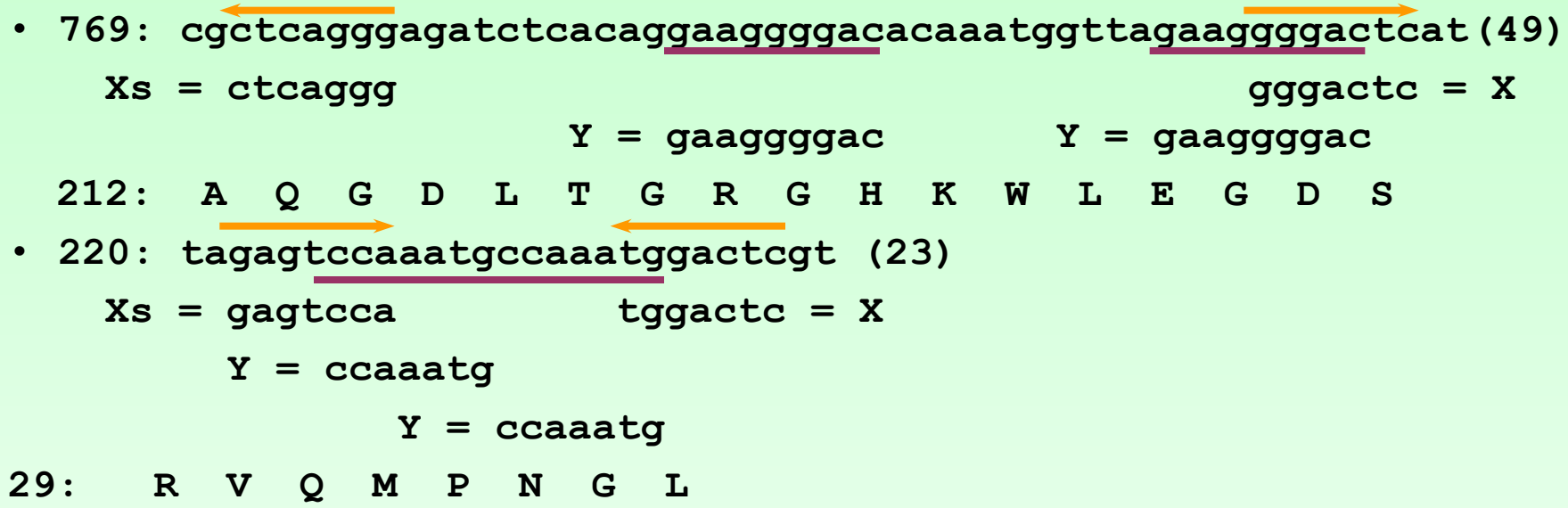
Комбинации разнотипных нерасширяемые повторов (прямой + инвертированный…) с ограничением на длины повторов: $l \geq R$ и длины «гэпов» $d \leq r$.

ДНК-последовательности + последовательности поведения животных

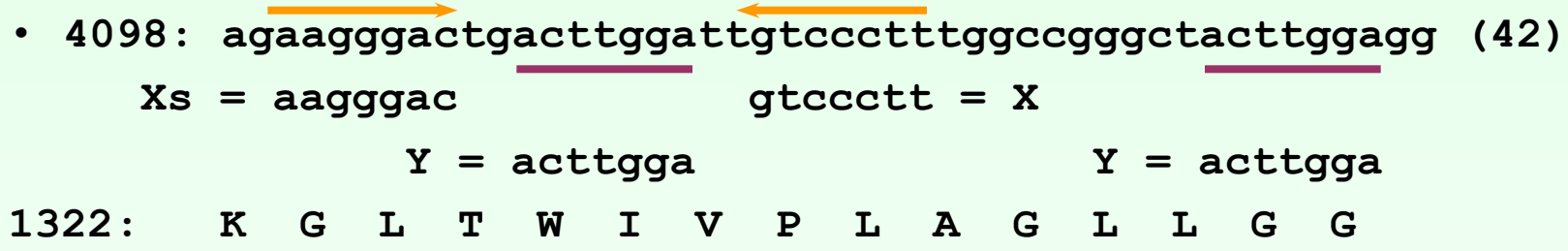


Комбинированные структуры в геноме ВКЭ

Dalnegorsk (и другие вирулентные)



Primorye-183 (и другие инаппарантные)



Заключение

Каждая «структурная единица» требует разработки «своего» алгоритма

СТРУКТУРНЫЕ АНАЛОГИИ В СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ РАЗЛИЧНОЙ ЯЗЫКОВОЙ ПРИРОДЫ

Благодарю за внимание!