Слайды для презентации к докладу

КЛАСТЕРИЗАЦИЯ ОРГАНИЗМОВ ПО ХАРАКТЕРИСТИКАМ СТРОЯ ИХ ДНК

Поздниченко Н.Н.

План доклада

- Средства исследования – веорятностно-информационные методы, анализ строя цепи
- Объект исследования – генетические тексты
- Цели исследования –кластеризация организмов, численное описание свойств строя генетического текста, выделение «естественных» информационных единиц в таких текстах

Методы анализа структуры символьных последовательностей

1. Методы, учитывающие только состав последовательности

2. Методы, учитывающие состав и косвенно – взаимное расположение элементов

3. Метод непосредственного исследования взаимного расположения элементов – анализ строя цепи

Строй цепи

N	M	M	N	V	M	V	V	N	M	N	V
A	L	L	A	B	L	B	B	A	L	A	B
F	G	G	F	H	G	H	H	F	G	F	H
1	2	2	1	3	2	3	3	1	2	1	3
1	-	-	1	-	-	-	-	1	-	1	-

Δ11 Δ12 Δ13 Δ14

где *Δji* интервал между *i*-м и *i*+1 вхождением элемента *j*-ой однородной цепи

Декомпозиция строя неоднородной знаковой цепи на однородные

V	*N*	*A*	*B*	*J*	*K*	*T*	*T*	*B*	*T*	*A*	*A*	*T*	*V*	*T*	*A*	*B*	знаковая цепь
1	2	3	4	5	6	7	7	4	7	3	3	7	1	7	3	4	строй цепи
1													1				однородные цепи и соответствующие им цепи интервалов
	2																
		3								3	3				3		
			4					4								4	
				5													
					6												
						7	7		7			7		7			

матрица интервалов				
0	0	0	13	4
0	0	0	0	16
0	8	1	4	2
0	0	5	8	1
0	0	0	0	13
0	0	0	0	12
1	2	3	2	3

Используемые в работе числовые характеристики строя

$$g = \log_2 \Delta_g = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log_2 \Delta_{ij}$$

$$r = \frac{\Delta_g}{D}$$

где m – мощность алфавита,

n_j – число вхождений j-го элемента,

D – макисмальный средний геометрический интервал

Некоторые понятия генетики

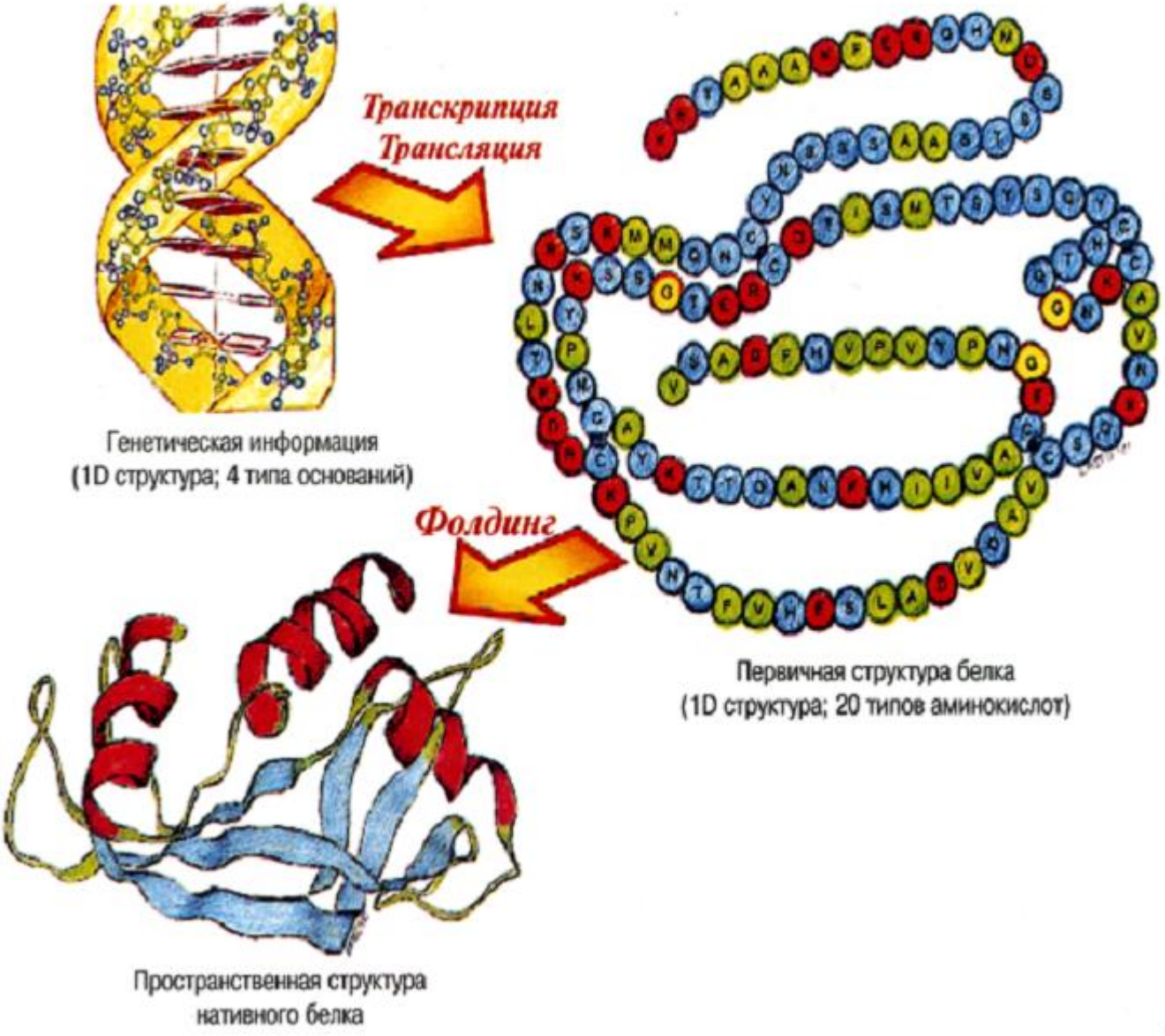
Нуклеотиды – являются элементарными компонентами нуклеиновых кислот, таких как ДНК и РНК.

Кодон , триплет – единица генетического кода, тройка нуклеотидных остатков в ДНК или РНК, кодирующая включение одной аминокислоты.

Аминокислоты – органические молекулы, являющиеся элементарными компонентами белков.

Белки – высокомолекулярные органические вещества, состоящие из соединённых в цепочку пептидной связью аминокислот.

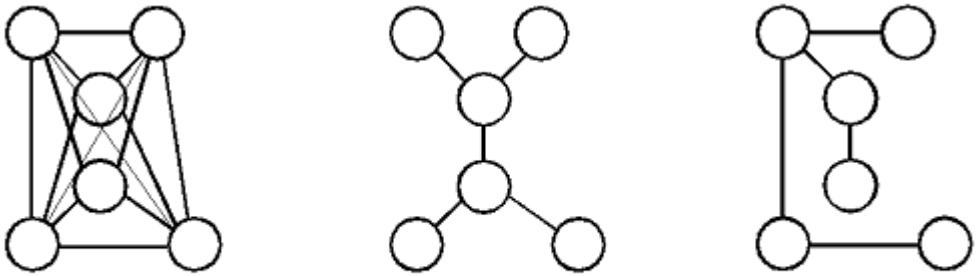
Структурные уровни организации белков



Классификация, основные понятия

- *Таксономия* — задача разбиения заданной выборки объектов (ситуаций) на подмножества, называемые таксонами (классами), так, чтобы каждый класс состоял из схожих объектов, а объекты разных таксонов существенно отличались.
- *Мерономия* – разделение объектов, позволяющее определить степень сходства между ними. Вычленение частей, образующих структуру объекта.
- *Классификация* – метод познания в котором результатом является представление знаний в виде некоторой классификационной схемы. В ней изучаемые объекты группируются в классы с помощью целесообразно выбранных признаков – оснований классификации.

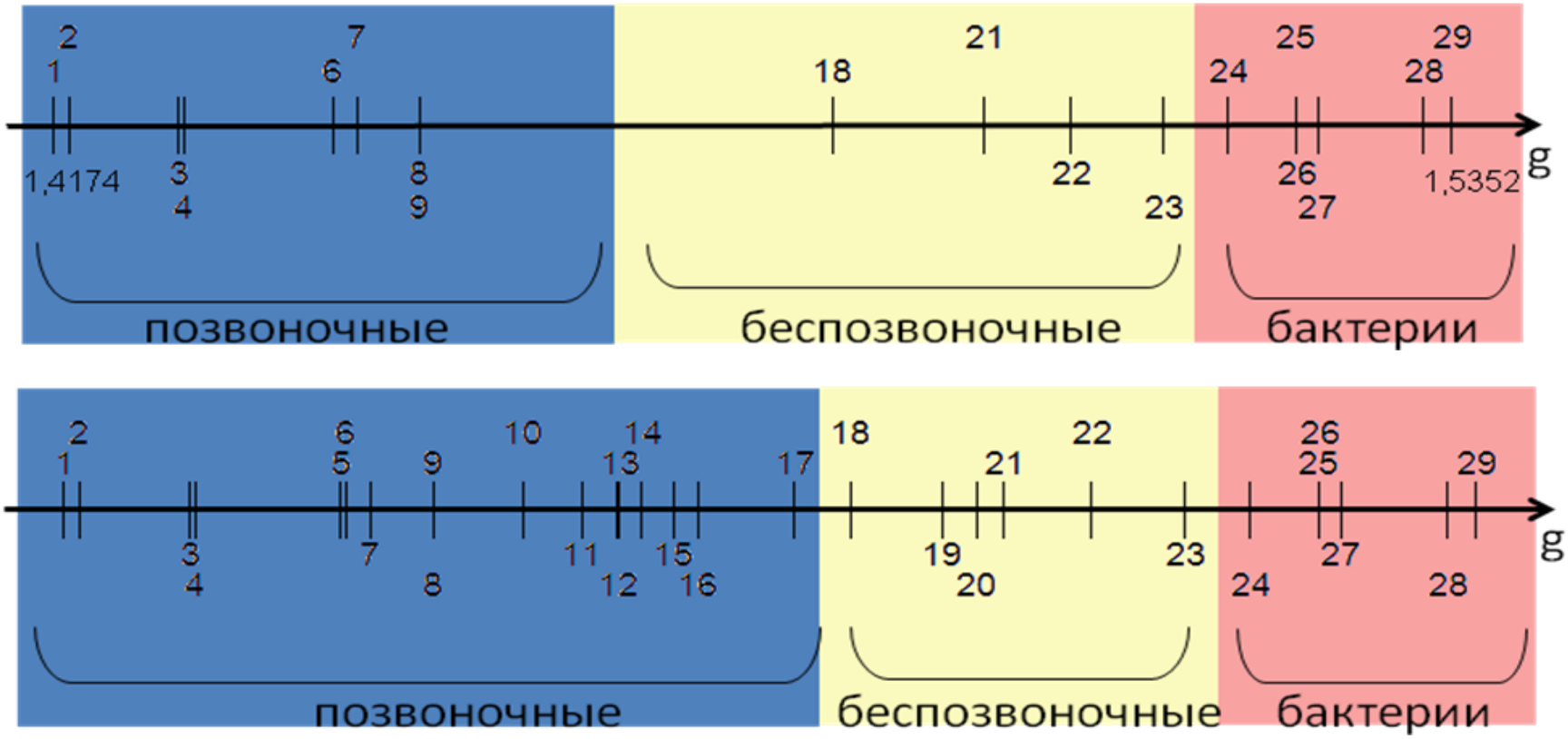
λ-KRAB



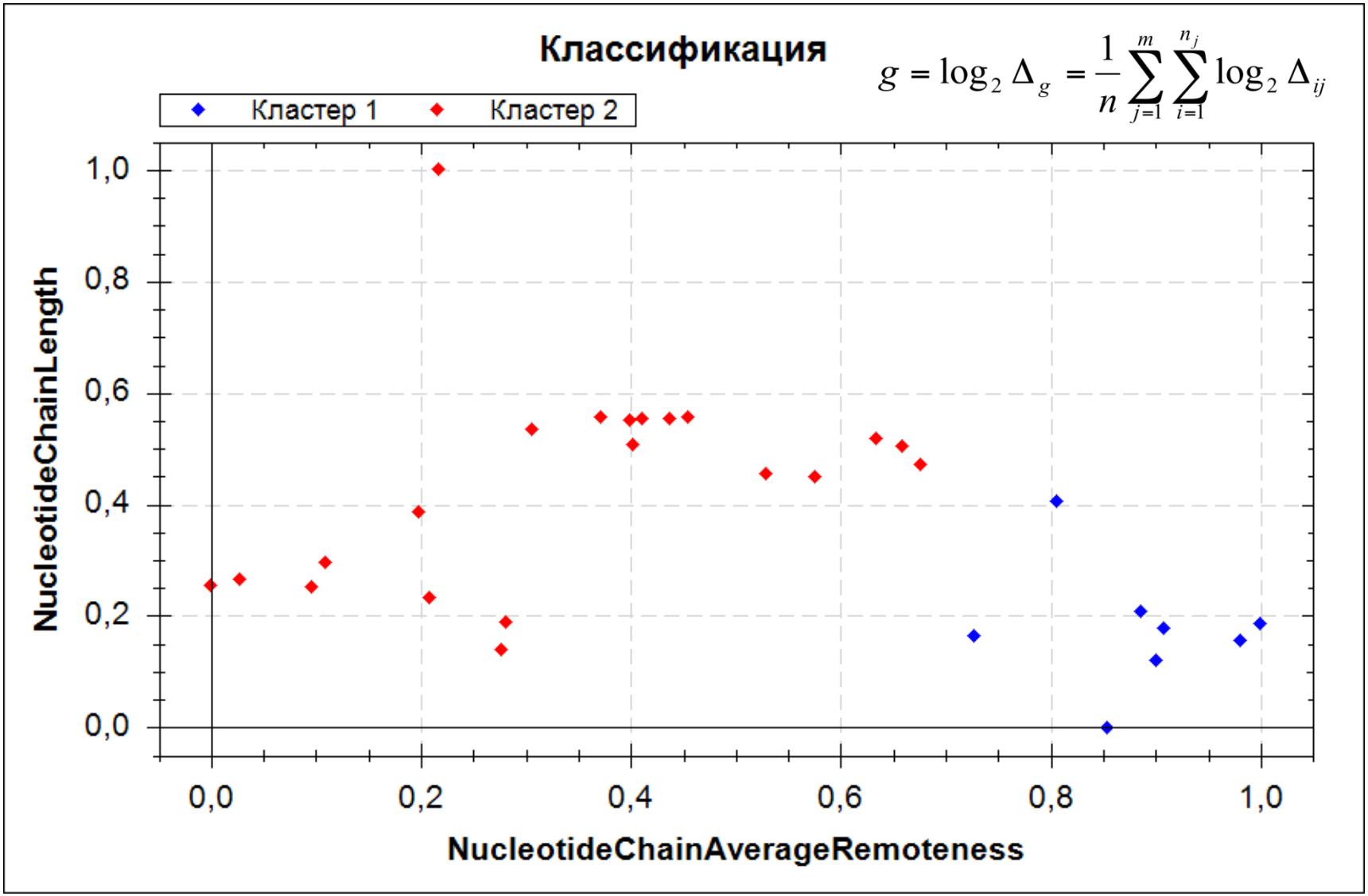
id	название	g
1	***M.musculus - мышь***	***1,417414581***
2	***C.crocodylus - Крокодил***	***1,418643679***
3	***C.familiaris - Собака***	***1,427959968***
4	***G.gallus - курица***	***1,428402795***
5	Sus scrofa - Кабан	1,440507111
6	***A.calva - рыба***	***1,440937027***
7	***H.s - человек***	***1,442946531***
8	***Th.thermophilus - микроорганизм***	***1,448283347***
9	***Th.thermarum - микроорганизм***	***1,448292643***
10	Gallus gallus - Банкивская джунглевая курица	1,455663685
11	Bos taurus 18S ribosomal RNA gene - Дикий бык	1,460651525
12	Erinaceus europaeus - Обыкновенный ёж	1,463618632
13	Homo sapiens - Человек разумный	1,463758914
14	Mus musculus - Домовая мышь	1,46560513
15	Cricetulus griseus - серый хомячок	1,468230019
16	Rattus norvegicus - Серая крыса	1,470291648
17	Crocodylus niloticus - Нильский крокодил	1,478396611
18	***I.persulcatus - Иксодовые клещи***	***1,483028363***
19	Zebrias zebra - Рыба	1,490709799
20	Kareius bicoloratus - двухцветная камбала	1,49358883
21	***O.moubata - клещи***	***1,495843469***
22	***P.humanus cap - блоха***	***1,503112618***
23	***M.domestica - муха***	***1,510952625***
24	***S.pyogenes - Стрептококк***	***1,51643989***
25	***B.anthracis - Сибирская язва***	***1,522106822***
26	***B.burgdorferi - боррелиоз***	***1,522174205***
27	***Candidatus N.m - бактерия***	***1,523941071***
28	***M.pneumoniae - атипичная пневмония***	***1,532855504***
29	***N.g - гонорея***	***1,535260296***

На числовой оси организмы расположены в соответствии со значениями характеристики *g* – средней удаленности элементов в нуклеотидной цепи.

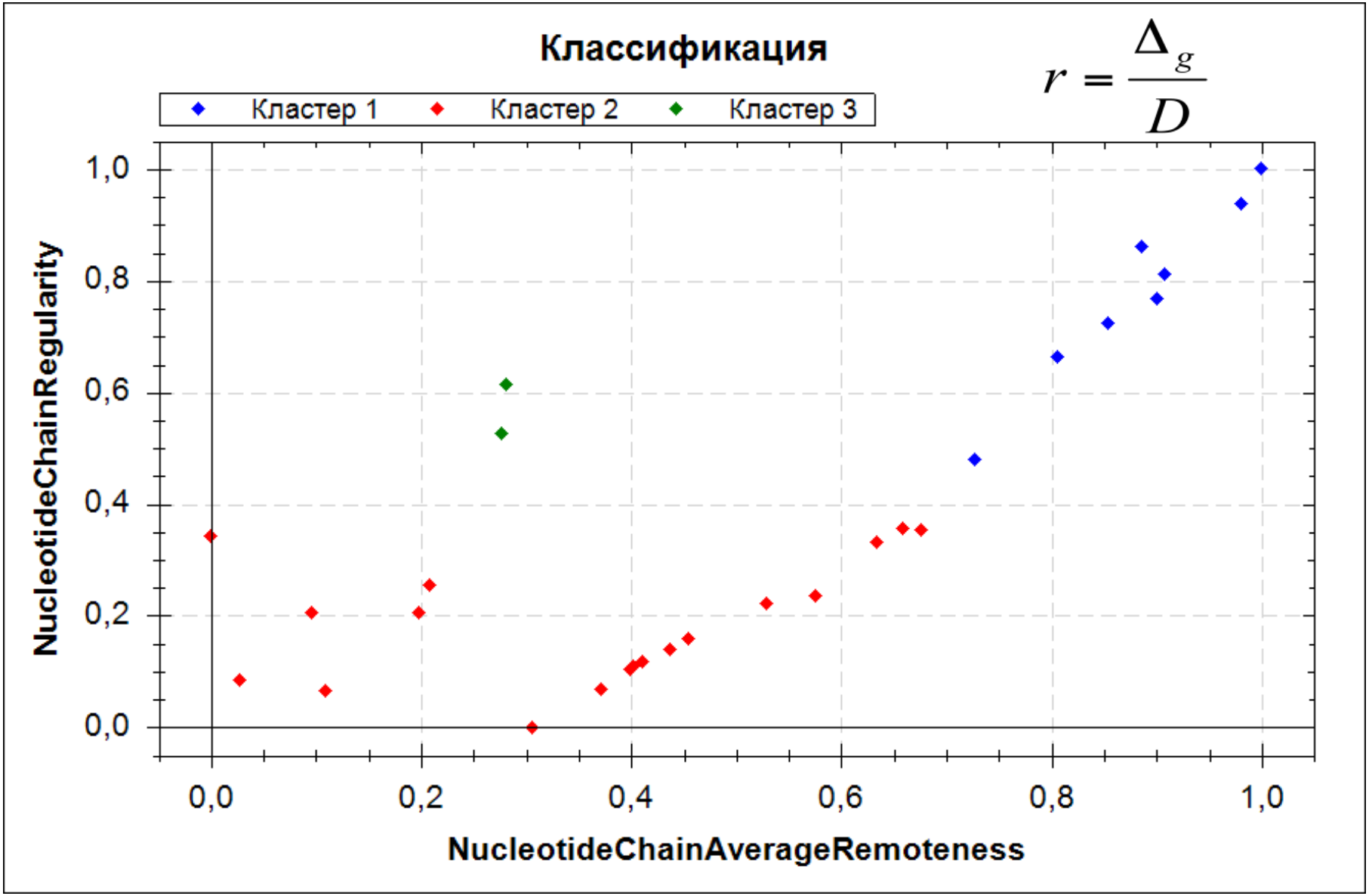
Эукариоты (позвоночные) - слева на оси, наименьшие значения g.
Прокариоты (бактерии) - справа на оси, наибольшие значения g.
Эукариоты (беспозвоночные) примыкают к бактериям.



id	Название	номер кластера
469	Candidatus N.m - бактерия	1
466	B.burgdorferi - боррелиоз	1
465	B.anthracis - Сибирская язва	1
479	S.pyogenes - Стрептококк	1
478	P.humanus cap - блоха	1
476	N.g гонорея	1
475	M.pneumoniae - Атипичная пневмония	1
473	M.domestica - муха	1
468	C.crocodylus - Крокодил	2
467	C.familiaris - Собака	2
464	A.calva - рыба	2
504	Zebrias zebra - Рыба	2
503	Sus scrofa - Кабан	2
502	Rattus norvegicus - Серая крыса	2
501	Mus musculus - Домовая мышь	2
500	Kareius bicoloratus - двухцветная камбала	2
499	Homo sapiens - Человек разумный	2
498	Gallus gallus - Банкивская джунглевая курица	2
497	Erinaceus europaeus - Обыкновенный ёж	2
496	Crocodylus niloticus - Нильский крокодил	2
495	Cricetulus griseus - серый хомячок	2
494	Bos taurus 18S ribosomal RNA gene - Дикий бык	2
481	Th.thermophilus - микроорганизм	2
480	Th.thermarum - микроорганизм	2
477	O.moubata - клещи	2
474	M.musculus - мышь	2
472	I.persulcatus - Иксодовые клещи	2
471	H.s - человек	2
470	G.gallus - курица	2


Классификация
$g = \log_2 \Delta_g = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log_2 \Delta_{ij}$
Кластер 1
Кластер 2
NucleotideChainLength
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
NucleotideChainAverageRemoteness

id	Название	номер кластера
469	Candidatus N.m – бактерия	1
466	B.burgdorferi – боррелиоз	1
465	B.anthracis - Сибирская язва	1
479	S.pyogenes – Стрептококк	1
478	P.humanus cap – блоха	1
476	N.g гонорея	1
475	M.pneumoniae - Атипичная пневмония	1
473	M.domestica – муха	1
468	C.crocodylus – Крокодил	2
467	C.familiaris – Собака	2
464	A.calva – рыба	2
504	Zebrias zebra – Рыба	2
502	Rattus norvegicus - Серая крыса	2
501	Mus musculus - Домовая мышь	2
500	Kareius bicoloratus - двухцветная камбала	2
499	Homo sapiens - Человек разумный	2
498	Gallus gallus - Банкивская джунглевая курица	2
497	Erinaceus europaeus - Обыкновенный ёж	2
496	Crocodylus niloticus - Нильский крокодил	2
495	Cricetulus griseus - серый хомячок	2
494	Bos taurus 18S ribosomal RNA gene - Дикий бык	2
477	O.moubata – клещи	2
474	M.musculus – мышь	2
472	I.persulcatus - Иксодовые клещи	2
471	H.s – человек	2
470	G.gallus – курица	2
481	Th.thermophilus - микроорганизм	3
480	Th.thermarum - микроорганизм	3


Классификация
$r = \frac{\Delta_g}{D}$
Кластер 1
Кластер 2
Кластер 3
NucleotideChainRegularity
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
NucleotideChainAverageRemoteness