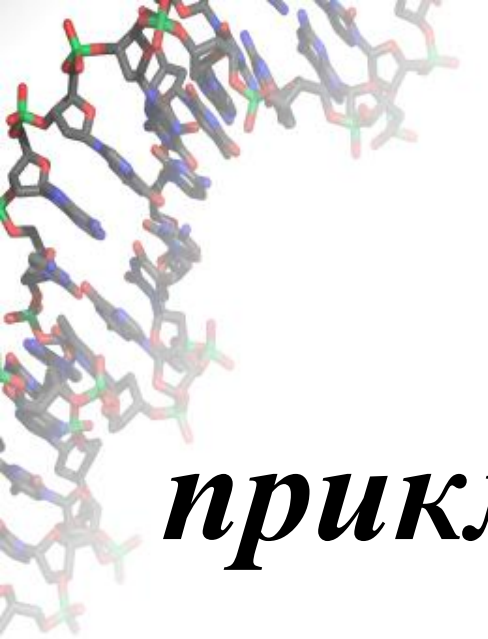


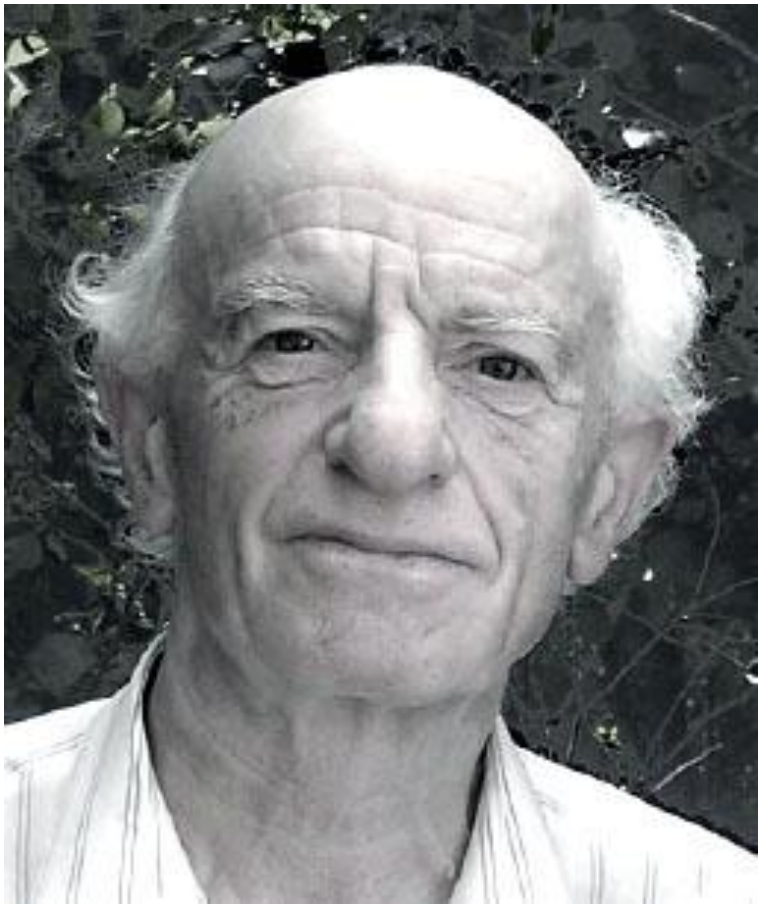
Лаборатория прикладной математики

М.А. Ройтберг

Пущино

27 апреля 2016

Лаборатория прикладной математики
Основатель лаборатории: проф. Э.Э. Шноль



(1928—2014)

Лаборатория прикладной математики
Сотрудники

№№	ФИО, степень	Должность	Возраст
1	Ройтберг М.А., д.ф.-м.н.	Зав. лаб.	
2	Фурлетова Е.И., к.ф.-м.н.	Научный сотрудник	до 35 лет
3	Яковлев В.В., к.ф.-м.н.	Научный сотрудник	до 35 лет
4	Астахова Т.В.	Научный сотрудник	
5	Ройтберг А.М.	М.н.с.	до 35 лет
6	Карпов А.В.	Вед.програм.	
7	Хачко Д.В.	Инженер	до 35 лет
8	Максимова З.И.	Инженер	
9	Карпова Н.Ф.	Инженер	
10	Акименко А.В.	Математ.	
11	Баулин Е.Ф.	асп., ИМПБ РАН	до 35 лет
12	Андреев Р.В.	асп., ИПМ РАН	до 35 лет
13	Поверенная И.А.	асп. ФББ МГУ	до 35 лет
14	Горев Д.И.	магистр., ФИВТ МФТИ	до 35 лет

Научное сотрудничество

- Институт белка РАН; ИППИ РАН;
- МГУ; МФТИ; НИУ ВШЭ

- INRIA/LIX team AMIB, Ecole Polytechnique, Palaiseau, France
- Institut Computational Biology LIRMM, Montpellier - France
- Ben Gurion University of the Negev, Ber-Sheva, Israel

- *NCBI, Bethesda, USA; Harvard University, Boston, USA; Georgia Tech, Atlanta USA;*
- *Lille University, Lille, France, etc,*

2011 – 2016 гг.

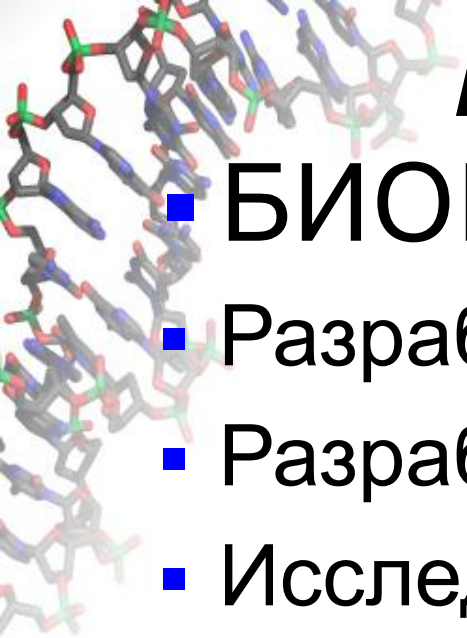
Публикации – 20

- Международные журналы (WoS, SCOPUS) – 10
- Российские журналы (ВАК) – 10

Участие в международных конференциях 11

Гранты

- РФФИ 09-04-01053 (2009-11) ; 12-04-00944 (2012-14); 14-01-93106 (2014-15) ; 14-04-32220 мол_а (2014-15); **16-04-01640** (**2016-18**);
- Государственный контракт «2011-1.4-514-008-009» (2011-12)



Направления исследований

- БИОИНФОРМАТИКА
- Разработка алгоритмов
- Разработка программ и сервисов
- Исследование биологических объектов
- ДРУГИЕ НАПРАВЛЕНИЯ
- Компьютерная лингвистика
- *Вычислительная математика*

Наиболее значимые результаты

- **подход к различным задачам как к задачам анализа графов и гиперграфов над полукольцами**

  Computation of biopolymers: a general approach to different problems
  AV Finkelstein, MA Roytberg - BioSystems,
  A unifying framework for seed sensitivity and its application to subset seeds
  G Kucherov, L Noé, M Roytberg - Journal of bioinformatics and …,

- **алгоритмы выравнивания биологических последовательностей для различных постановок задач**

  A hierarchical approach to aligning collinear regions of genomes
  MA Roytberg, AY Ogurtsov, SA Shabalina… - …, Bioinformatics

- ***Increasing the accuracy of global alignment of amino acid sequences by constructing a set of alignment candidates***
  VV Yakovlev, MA Roytberg - Biophysics

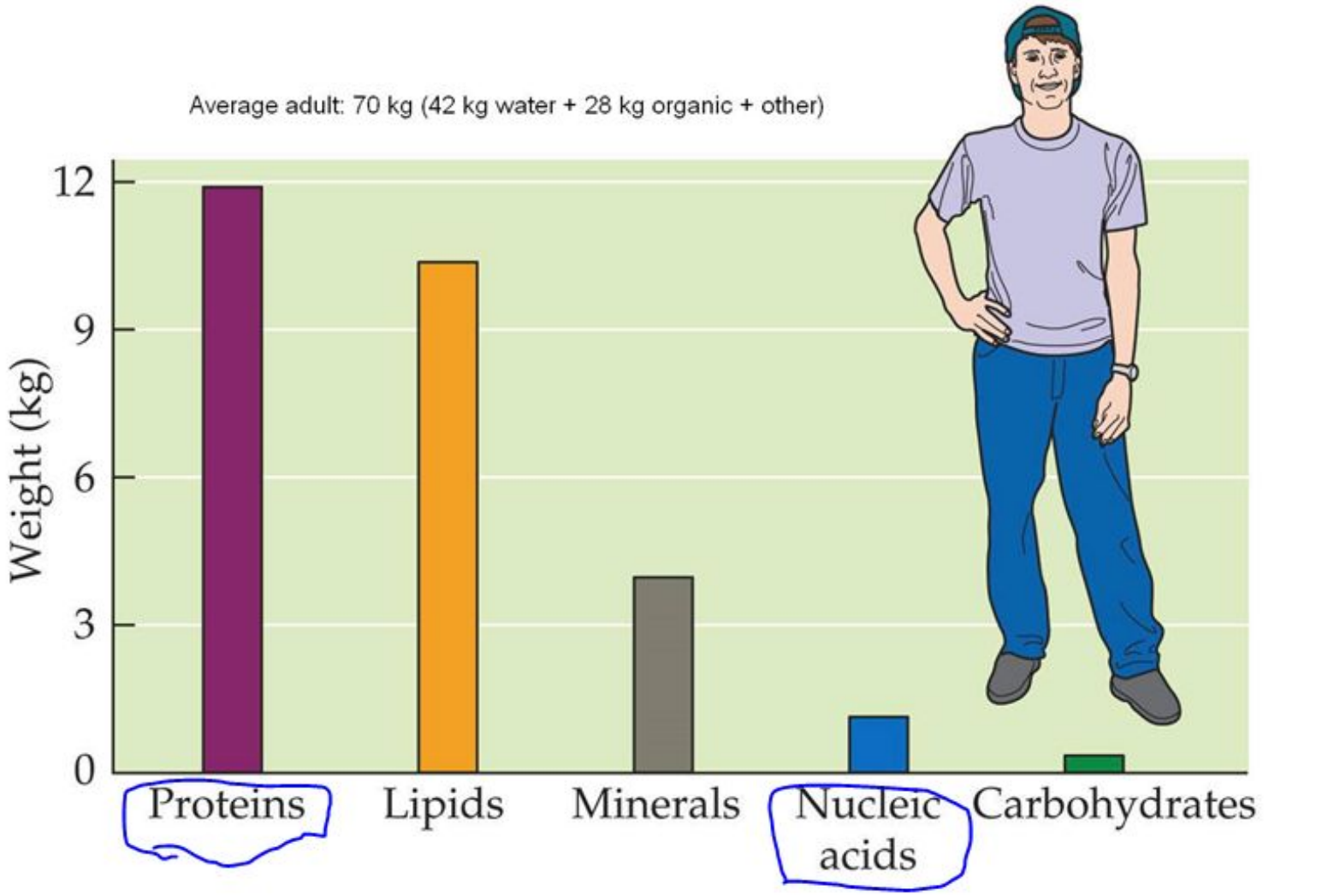
- **алгоритмы построения и анализа затравок для поиска локальных сходств** (совм.с гр. Г.Кучерова)

  Multiseed lossless filtration *G Kucherov, L Noé, M Roytberg - IEEE/ACM Transactions on …,* ***On subset seeds for protein alignment*** *E Furletova, E Szczurek, G Kucherov -*

Активные направления исследований

- Анализ экзон-интронных структур
- Анализ пространственных структур РНК
- Анализ мотивов в биологических последовательностях
- Сравнительный анализ биологических последовательностей: геномные выравнивания.
- Компьютерная лингвистика (анализ бриджинга и анафоры)

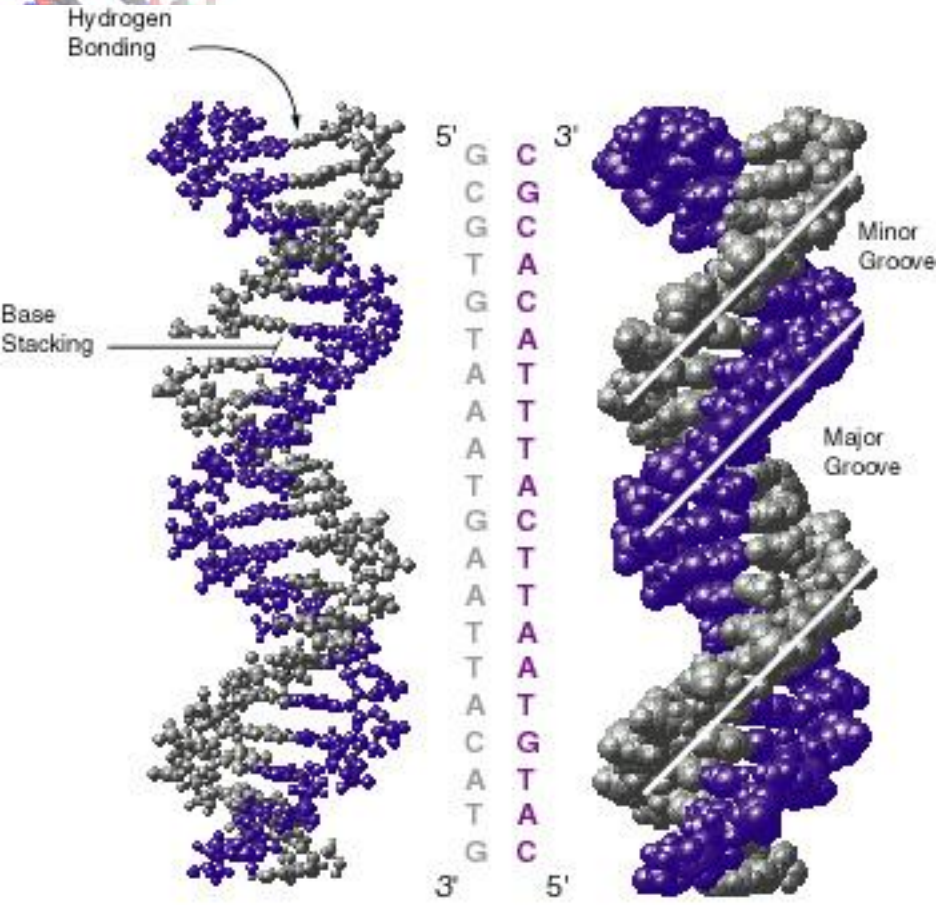
Биополимеры: нуклеиновые кислоты (ДНК, РНК); белки

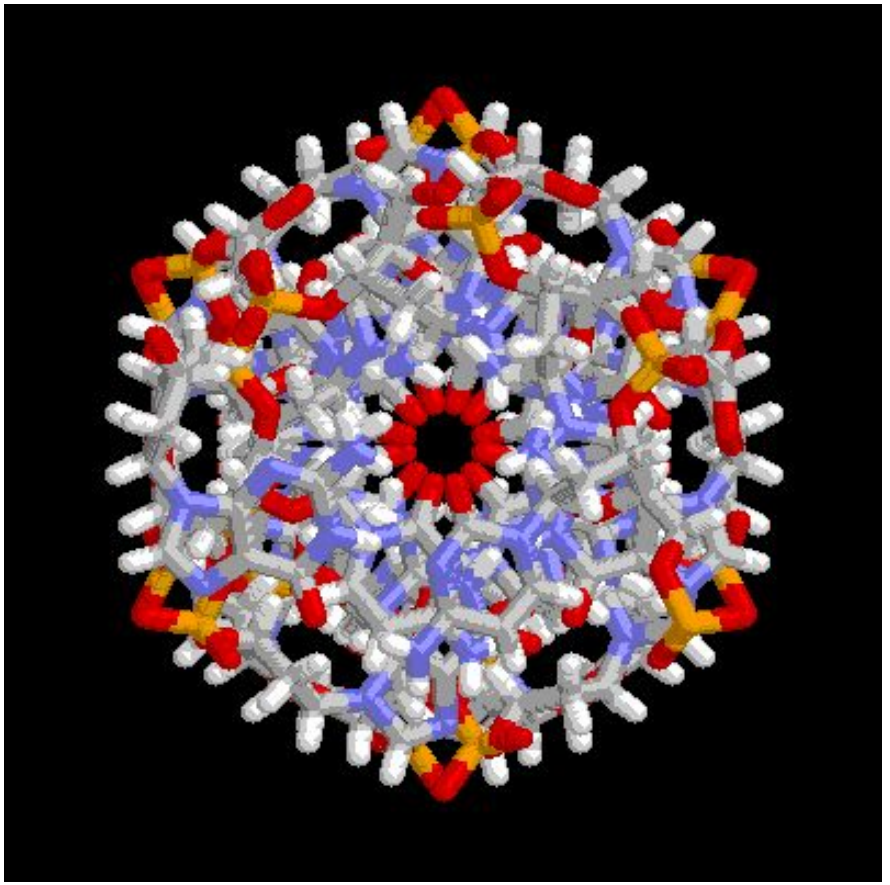


ANIMAL PHYSIOLOGY, Figure 4.1 © 2004 Sinauer Associates, Inc.

ДНК: 2 нити; L ~ 10^5 – 10^9

нуклеотиды (4)

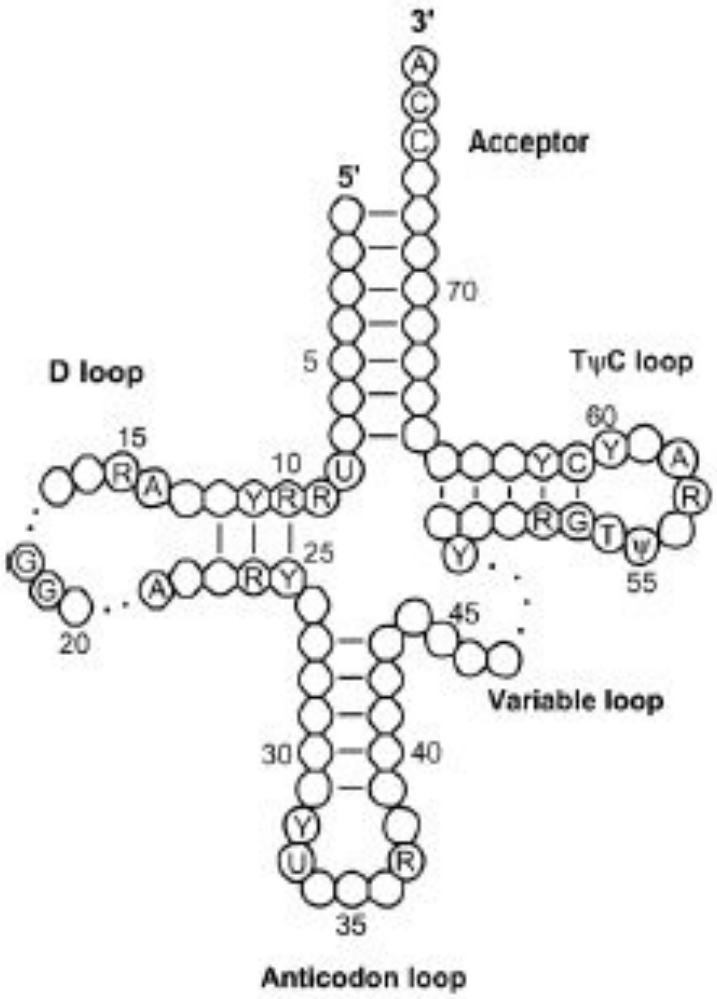


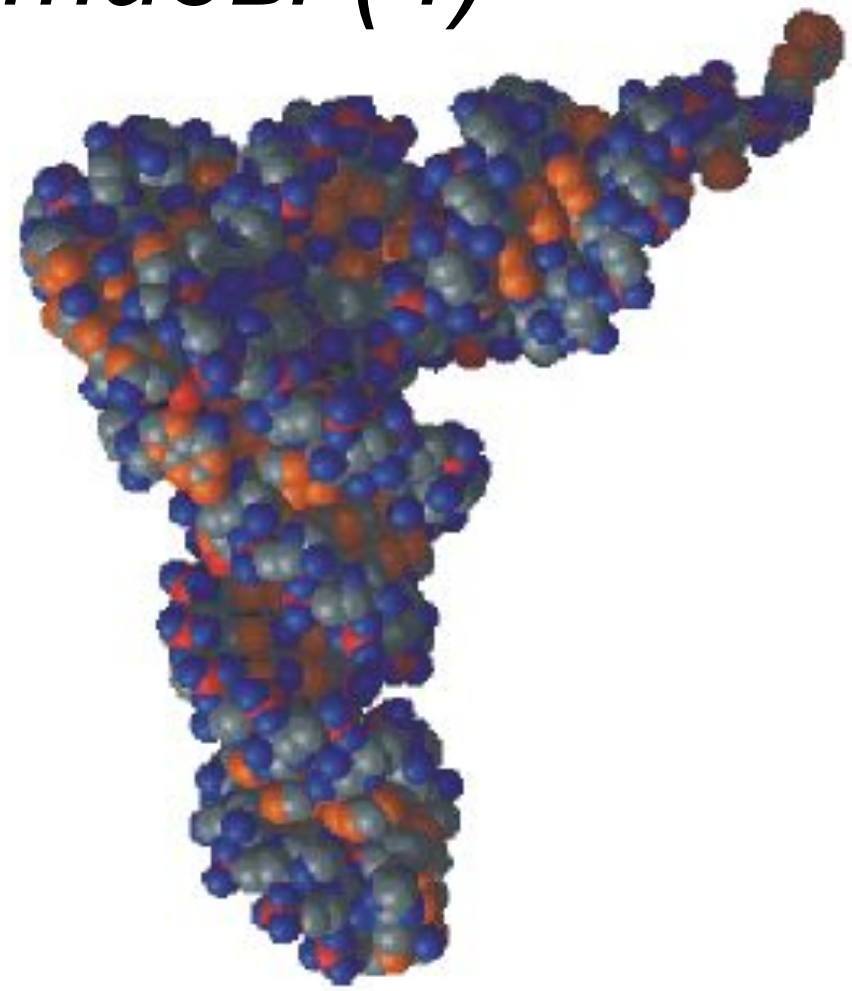


Miesfeld / Applied Molecular Genetics
Fig. 01.03. DNA Structure.

РНК: 1 нить, L ~ 10^2 = 10

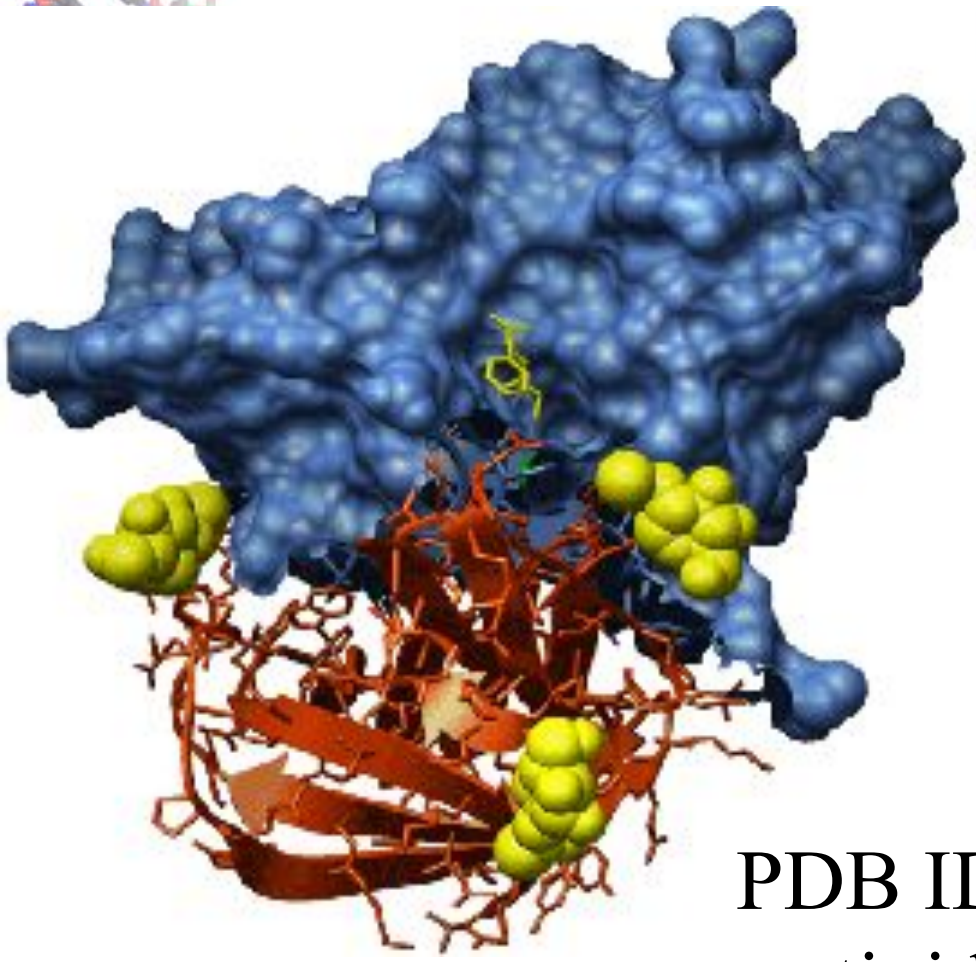
нуклеотиды (4)

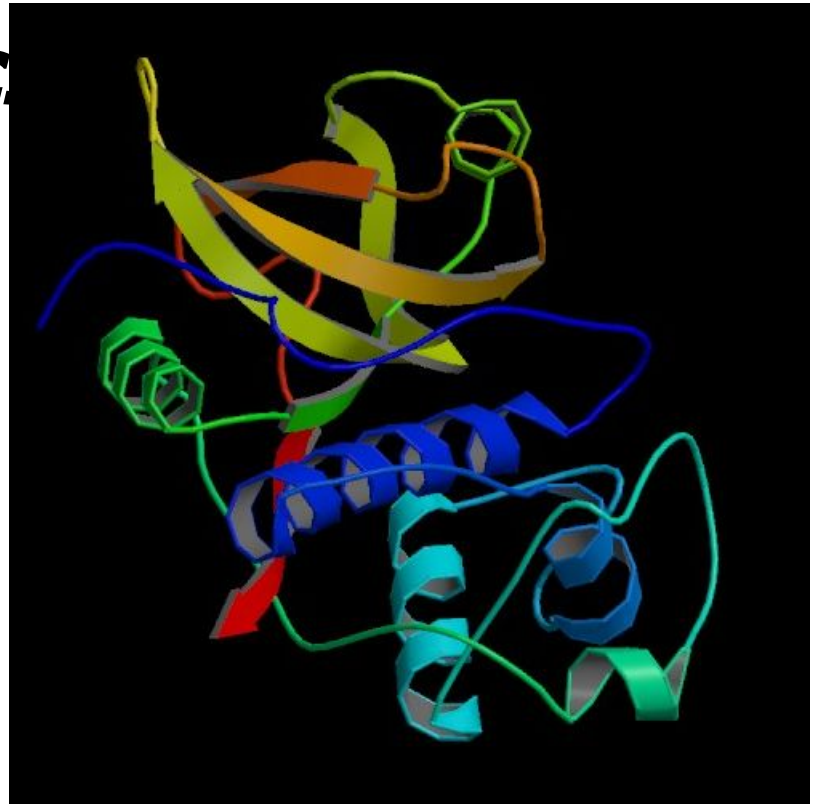




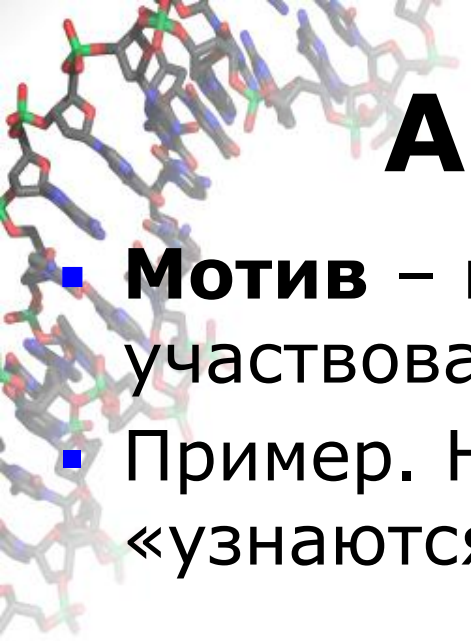
From Paul Higgs

– 10^3





PDB ID: **2act** The structure of actinidin at 1.7 Ångstroms



A. Анализ мотивов

- **Мотив** – набор олигомеров, которые могут участвовать в определенном процессе.
- Пример. Набор фрагментов ДНК, которые «узнаются» определенным белком.

Анализ мотивов

- **Мотив** – набор олигомеров, которые могут участвовать в определенном процессе.
- Пример. Набор фрагментов ДНК, которые «узнаются» определенным белком.
- ***Слово*** – последовательность букв в алфавите *A*

Алфавит нуклеотидов: {A,C,G,T}

Алфавит аминокислот*:*{A,G,P,R,W,C,D,E,H,Q,S,T,V,L,M,F,K,N,Y,I}

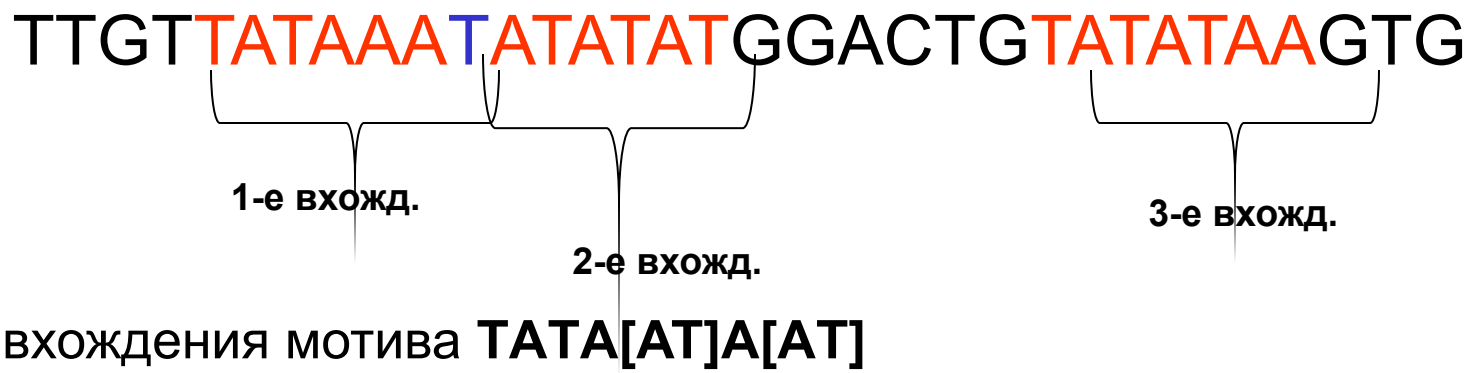
- ***Длина слова* –** количество букв в слове
- ***Мотив*** – набор слов одинаковой длины

***Примеры мотивов*:**

- **TATA[AT]A[AT]** – TATA-бокс (4 слова длины 7)
- **ANNNNCATTA** – сайт связывания белке Antp (Drosophila) (256 слов длины 10)

Вхождение мотива

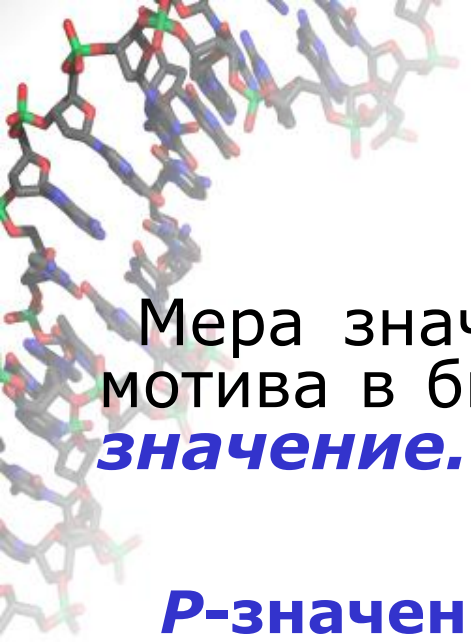
- Поиск сайтов связывания факторов регуляции транскрипции[1]
- Поиск функциональных участков в белках, например, неупорядоченных (не имеющих фиксированной пространственной структуры) участков[2].



вхождения мотива **TATA[AT]A[AT]**

1 Stormo G.D. **DNA binding sites: representation and discovery** // Bioinformatics. 16(1):16–23, 2000

2 Lobanov M.Y., Furletova E.I., Bogatyreva N.C., Roytberg M.A., Galzitskaya O.V. **Library of disordered patterns in 3D protein structures**// PLoS Computational Biology. 6(10). 2010.



P-значение

Мера значимости найденной группы вхождений мотива в биологической последовательности – ***P-значение.***

***P*-значение** ($P(r,n)$) – вероятность встретить мотив H не менее r раз в случайной последовательности длины n

Цель исследования – разработать эффективный алгоритм для нахождения P-значения

Алгоритмы

AhoPro[1], SufPref[2]– алгоритмы для вычисления точного P-значения

OWA[3] – алгоритм для нахождения приближенного P-значения

1. Boeva V, Clément J, Régnier M, Roytberg MA, Makeev VJ **Exact p-value calculation for heterotypic clusters of regulatory motifs and its application in computational annotation of cis-regulatory modules** // Algorithms for molecular biology, 2:13,2007. **http://bioinform.genetika.ru/AhoPro/**

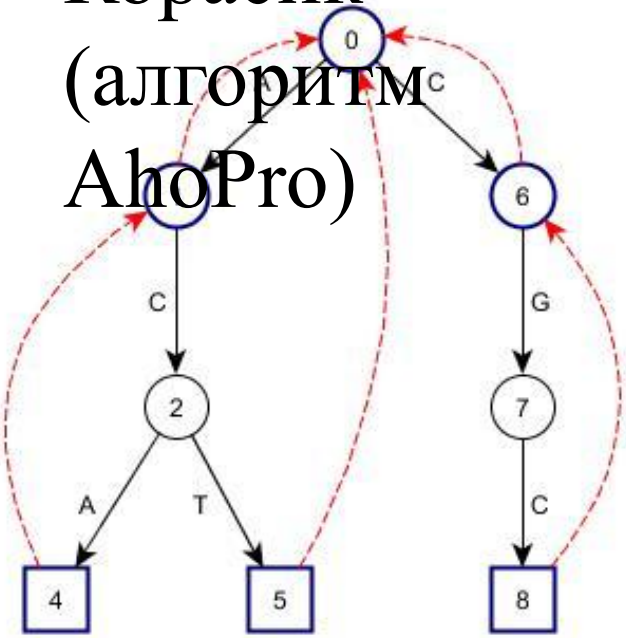
2. Regnier M., Furletova E.I., Yakovlev V.V., Roytberg M.A. **Analysis of pattern overlaps and exact computation of P-values of pattern occurrences numbers: case of Hidden Markov Models.** // Algorithms for molecular biology, 9(1):25,2014. **http://server2.lpm.org.ru/bio/online/sf/**

3. Furletova EI, Holub J., Regnier M. **Minimized Compact Automaton for Clumps over Degenerate Patterns** // SeqBio'15, 2015

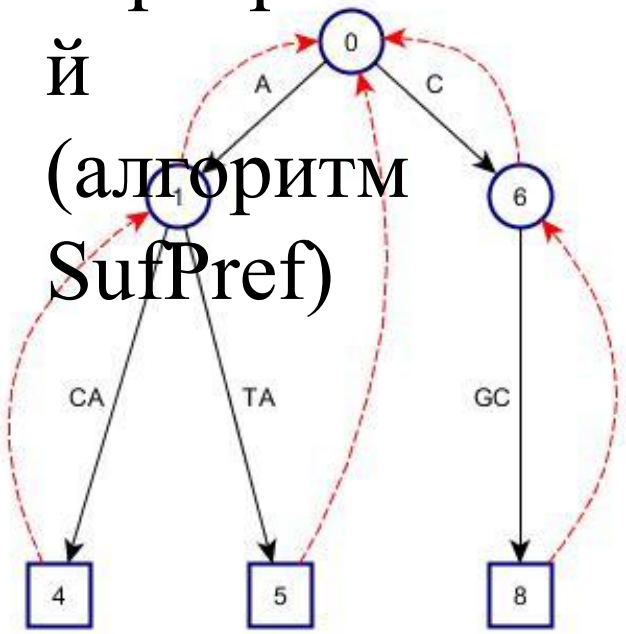
Все три алгоритма для нахождения P-значения используют специальные графы

Граф перекрытий (мотив ACA, ACT, CGC)

Граф Ахо-Корасик (алгоритм AhoPro)



Граф перекрытий (алгоритм SufPref)

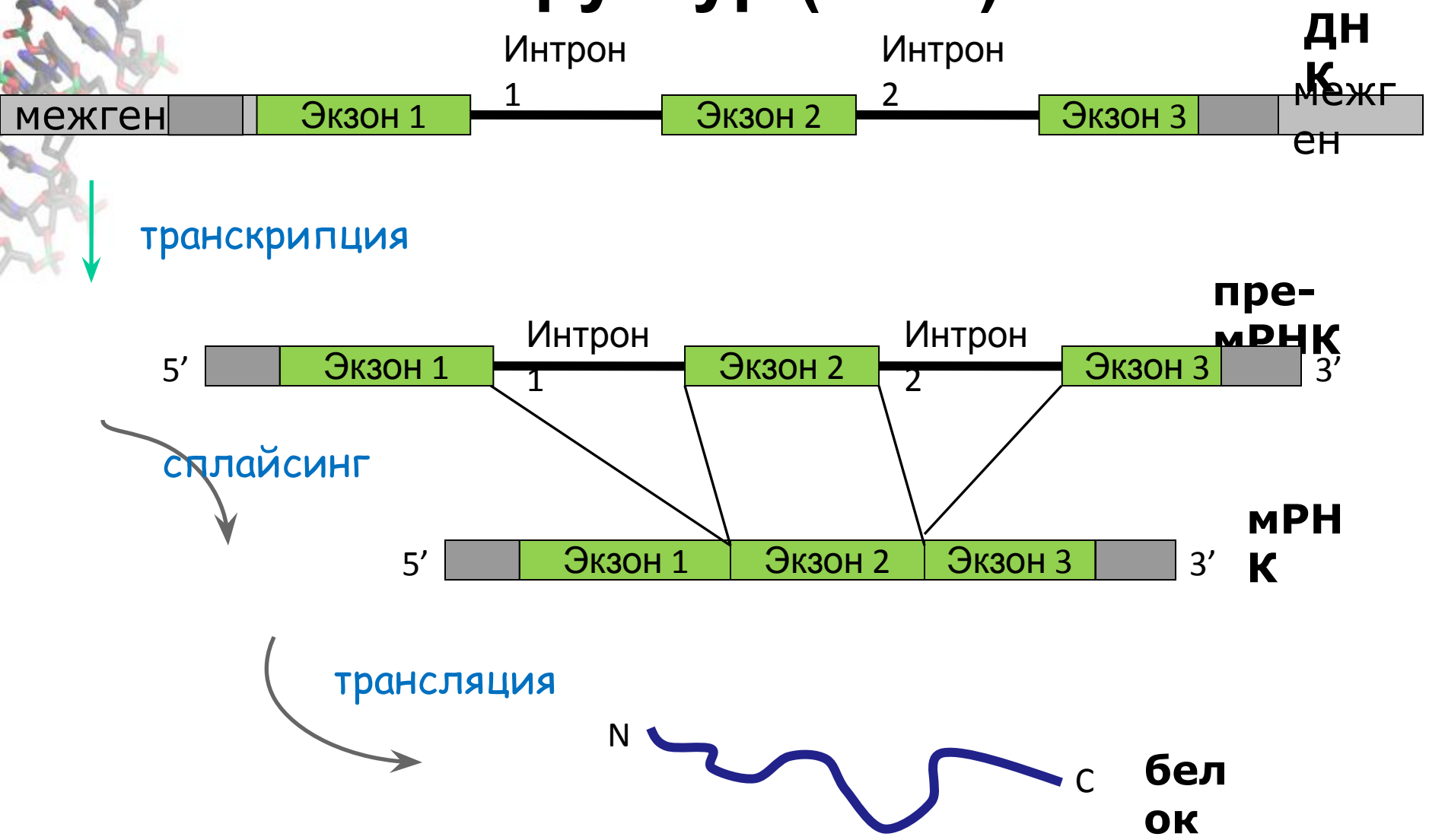


Паттерн, заданный матрицей позиционных весов с порогом 4.3, описывающий сайты фактора ANDR в геноме человека, 4270349 слов длины 16.

SufPref время: 12.71 секунд, память: 691.58 мегабайт

AhoPro: время: 351.6 секунд, память: 1868,18 мегабайт

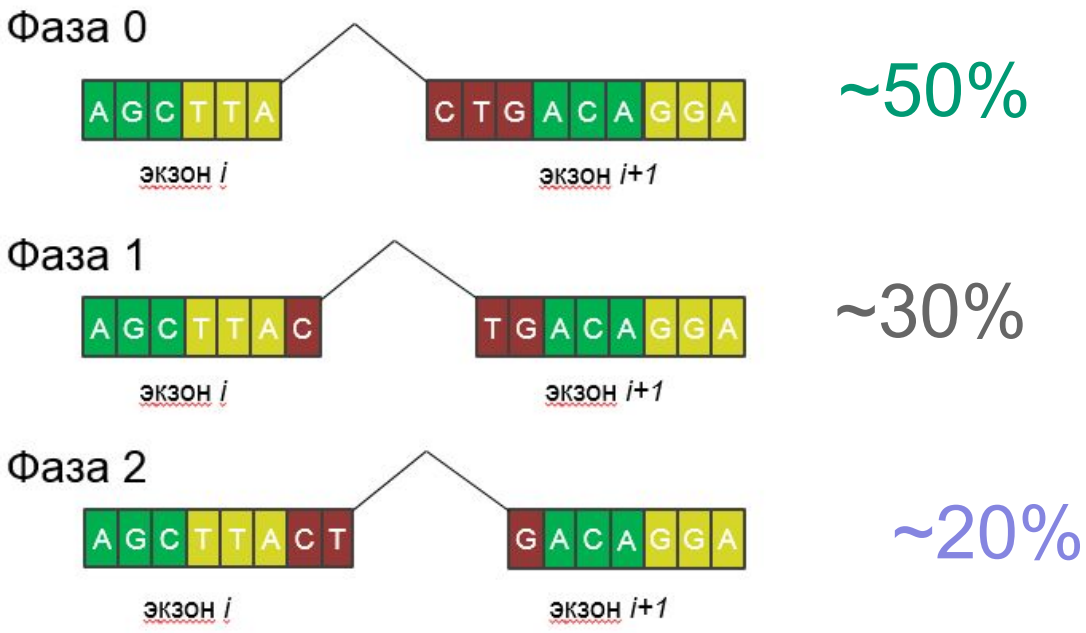
Б. Анализ экзон-интронных структур (ЭИС)



Интроны

- Крайне неконсервативная последовательность
- Есть почти у всех эукариот
- Плотность интронов в геноме варьируется у разных организмов: организмы разделяют на интрон-богатые и интрон-бедные.

Фаза интронов



~50%

~30%

~20%

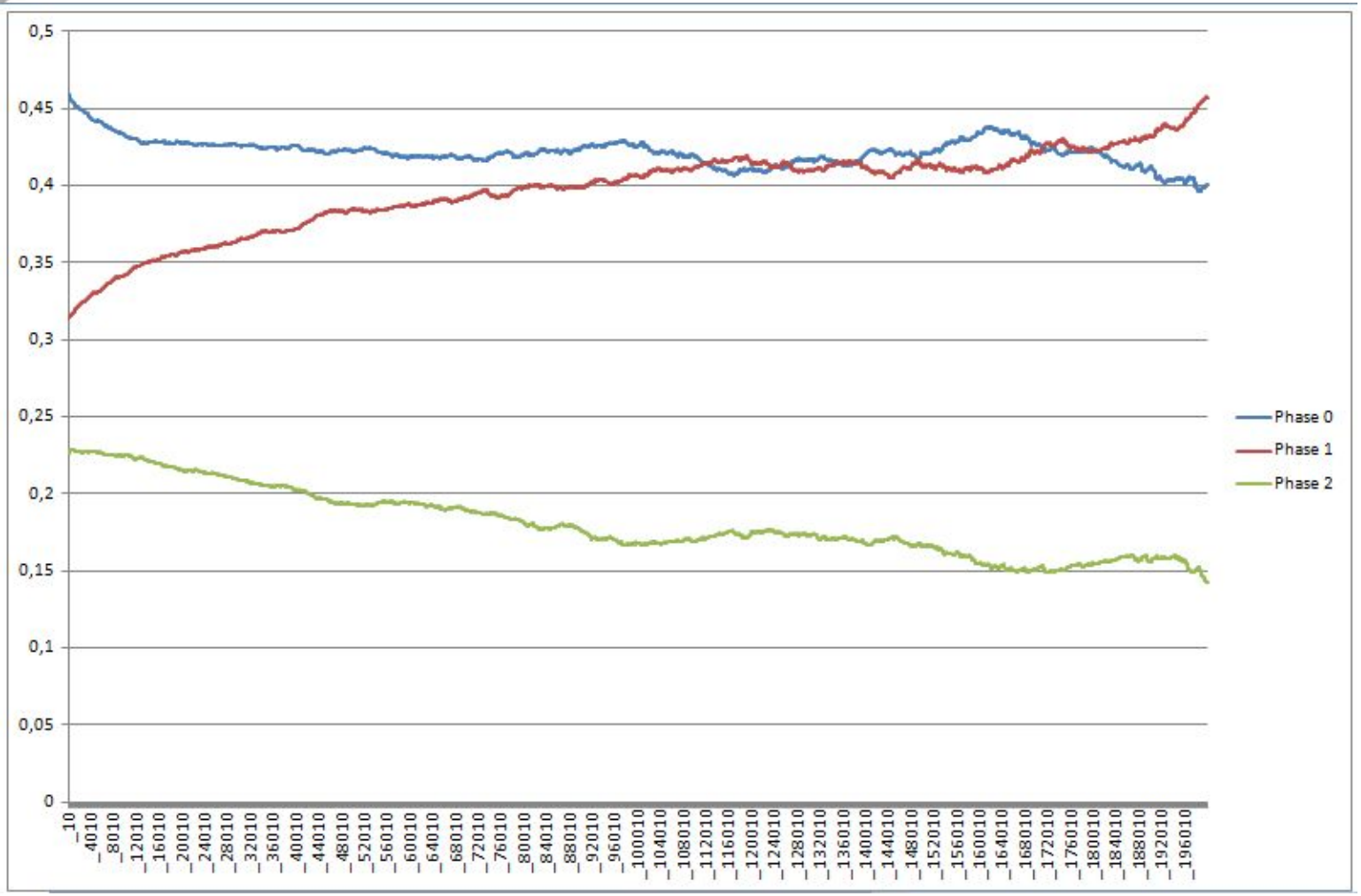
Вопросы

1. Какие существуют закономерности, связанные с фазой и/или длиной интрона?
2. Какие интроны считать гомологичными?
3. Как часто происходит изменение фазы («слайдинг»)?

Алгоритмические и программные задачи

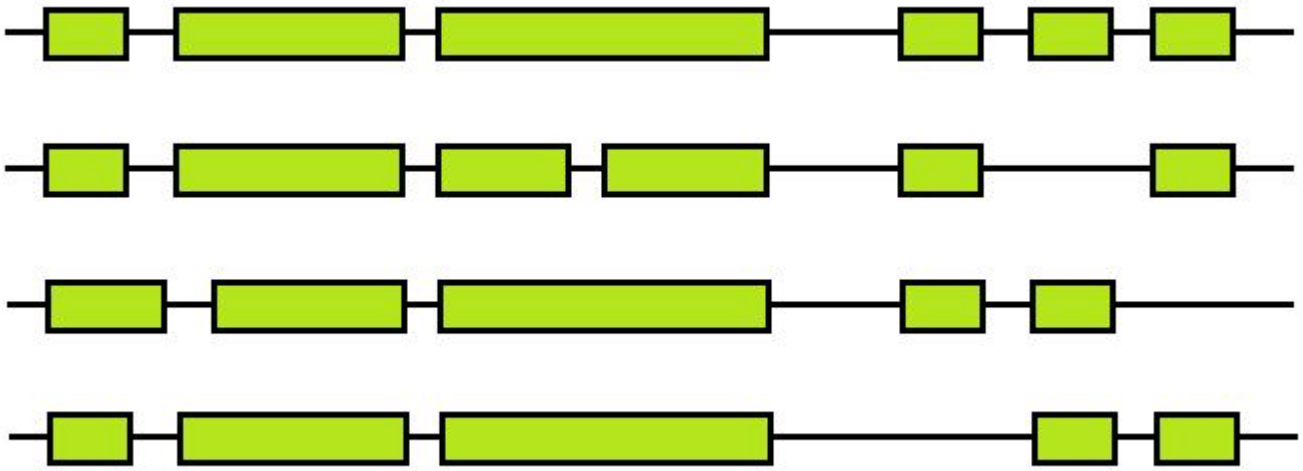
1. Множественное выравнивание экзон-интронных структур: уточнение постановки задачи, алгоритм, программа.
2. База данных экзон-интронных структур генов из эукариотических организмов.

1. Анализ длин и фаз интронов и поиск закономерностей



Фаза vs длина интрона

2. Разработка программы множественного ЭИС выравнивания



Входные данные:

- координаты экзонов
- множественное выравнивание белковых продуктов генов

Выравнивание ЭИС

Предварительные выравнивания для 4 генов были построены полуавтоматически и проверены вручную

организм		гр. 1		гр.2		гр. 3	гр. 4	гр. 5	гр. 6		гр. 7		гр. 8	гр. 9	гр. 10	гр. 11	гр. 12	гр. 13	гр. 14	гр. 15	гр. 16	гр. 17
хрящевые рыбы	1			2		3	4	5	6		7		8	9	10	11	12	13	14	15	16	17
костистые рыбы		1		2	3	4	5	6	7	8	9		10	11	12	13	14	15	16	17	18	
бесхвостые		1		2		3																
крокодилы	1												2	3	4	5	6	7	8	9	10	11
птицы	1			2		3	4	5	6		7		8	9	10	11	12	13	14	15	16	17
черепахи		1		2		3	4	5	6		7		8	9	10	11	12	13	14	15	16	17
кистеперовые рыбы		1		2		3	4	5	6		7		8	9	10	11	12	13	14	15	16	17
чешуйчатые	1							2	3		4		5	6	7	8	9	10	11	12	13	14
афротерии								1	2		3	4	5	6	7	8	9	10	11	12	13	14
грызунообразные		1		2		3	4	5	6		7		8	9	10	11		12	13	14	15	16
приматы	1			2		3	4	5	6		7		8	9	10	11	12	13	14	15	16	17
Хищные				1		2	3	4	5		6		7	8	9	10	11	12	13	14	15	16
Китопарнокопытные			1	2		3	4	5	6		7		8	9	10	11	12	13	14	15	16	17
рукокрылые	1						2	3	4		5		6	7	8	9	10	11	12	13	14	15
насекомоядные		1		2		3	4	5	6		7		8	9	10	11	12	13	14	15	16	17
непарнокопытные	1			2		3	4	5	6		7		8	9	10	11	12	13	14	15	16	17
броненосцы		1		2		3	4	5	6		7		8	9	10	11	12	13	14	15	16	17
сумчатые													1	2	3	4	5	6	7	8	9	
яйцекладущие								1	2		3		4	5	6	7	8	9	10	11	12	13

Ген MDGA2
Гомологи из 135 организмов

выровнено
выравнивается с тем же экзоном, что и сосед
частично выровнено
не выровнено

3. База данных ЭИС (EIS DB)

Исходные данные:
полногеномные сборки
из БД RefSeq
13 таблиц:
- Структурные
- Таксономические
- Таблицы экзон-интронной структуры
- Вспомогательные таблицы

Позвоночные	птицы	4
Позвоночные	рыбы	8
Позвоночные	млекопитающие	23
Позвоночные	рептилии	2
Беспозвоночные	насекомые	9
Беспозвоночные	круглые черви	1
Беспозвоночные	другие животные	1
Растения	высшие растения	26
Растения	зеленые водросли	4
Грибы	аскомицеты	28
беспозвоночные	базидомицеты	3
Грибы	другие грибы	4
Простейшие	споровики	11
Простейшие	кинетопластиды	8
Простейшие	другие простейшие	3
Другие	другие	3

Идет загрузка 138 геномов

Существующие базы данных не поддерживаются или узкоспецифичны.
В 2015 году вышла новая база JuncDB для изучения эволюции интронов.
Однако она не содержит последовательности интронов.

3. База данных ЭИС (EIS DB)

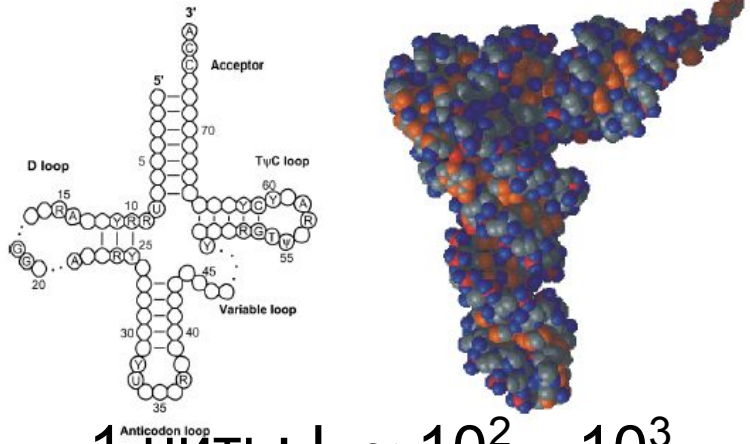
Исходные данные:
полногеномные сборки
из БД RefSeq
13 таблиц:
- Структурные
- Таксономические
- Таблицы экзон-интронной структуры
- Вспомогательные таблицы

Позвоночные	птицы	4
Позвоночные	рыбы	8
Позвоночные	млекопитающие	23
Позвоночные	рептилии	2
Беспозвоночные	насекомые	9
Беспозвоночные	круглые черви	1
Беспозвоночные	другие животные	1
Растения	высшие растения	26
Растения	зеленые водоросли	4
Грибы	аскомицеты	28
беспозвоночные	базидомицеты	3
Грибы	другие грибы	4
Простейшие	споровики	11
Простейшие	кинетопластиды	8
Простейшие	другие простейшие	3
Другие	другие	3

Идет загрузка 138 геномов

Существующие базы данных не поддерживаются или узкоспецифичны.
В 2015 году вышла новая база JuncDB для изучения эволюции интронов, однако она не содержит интронные последовательности.

Астахова Т.В., Ройтберг М.А. Цитович И.И., Яковлев В.В. **Закономерности, связанные с распределением длин интронов.** Математическая биология и биоинформатика. 2014. Т.9. №2. с. 482-490.



3. Пространственная структура РНК

1 нить; L ~ 10^2 – 10^3

нуклеотиды (4)

- Пространственная структура РНК играет ключевую роль в различных внутриклеточных процессах, протекающих с участием РНК.
- Типичные мотивы пространственной структуры РНК не систематизированы и, как следствие, недостаточно изучены.
- Прежде всего это относится к структурным мотивам, образованным с участием не Уотсон-Криковских взаимодействий и содержащим «псевдоузлы» («неканонические структуры»)

Источник данных:

- Экспериментально определенные пространственные структуры РНК

RCSB PDB
PROTEIN DATA BANK

Organism
- Homo sapiens (273)
- Escherichia coli (270)
- Thermus thermophilus HB8 (205)
- synthetic construct (177)
- Saccharomyces cerevisiae (166)
- Escherichia coli K-12 (134)
- Thermus thermophilus (104)
- Other (877)

Taxonomy
- Unassigned (868)
- Bacteria (867)
- Eukaryota (664)
- Viruses (303)
- Archaea (204)
- Other (185)

Experimental Method
- X-ray (2035)
- Solution NMR (582)
- Electron Microscopy (210)
- Fiber Diffraction (5)
- Hybrid (3)
- Other (1)
- Solid-State NMR (1)

X-ray Resolution
- less than 1.5 Å (85)
- 1.5 - 2.0 Å (219)
- 2.0 - 2.5 Å (378)
- 2.5 - 3.0 Å (598)
- 3.0 and more Å (755)
- more choices...

Release Date
- before 2000 (231)
- 2000 - 2005 (464)
- 2005 - 2010 (802)
- 2010 - 2015 (1194)
- 2015 - today (146)
- this month (20)
- more choices...

Polymer Type
- Mixed (1751)
- RNA (1086)

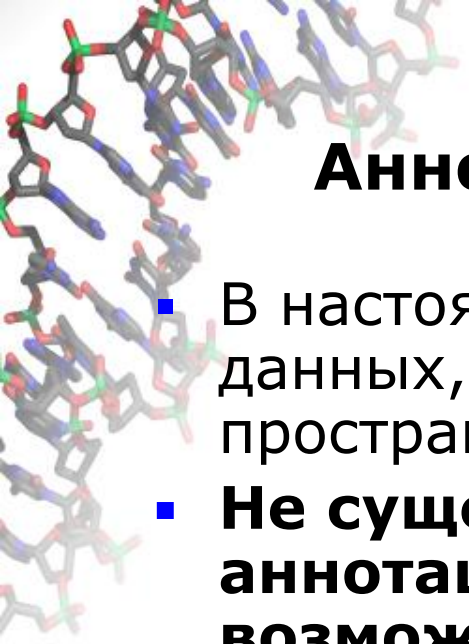
Enzyme Classification
- 2: Transferases (272)
- 3: Hydrolases (233)
- 6: Ligases (93)
- 5: Isomerases (26)
- 4: Lyases (10)
- 1: Oxidoreductases (4)

SCOP Classification
- Alpha and beta proteins (a+b) (286)
- All alpha proteins (227)
- All beta proteins (206)
- Alpha and beta proteins (a/b) (196)
- Small proteins (126)
- Peptides (91)
- Low resolution protein structures (81)
- Other (51)

Protein Symmetry
- Asymmetric (1129)
- Cyclic (204)
- Icosahedral (49)
- Dihedral (25)
- Helical (15)
- more choices...

Protein Stoichiometry
- Monomer (654)
- Heteromer (471)
- Homomer (297)
- more choices...

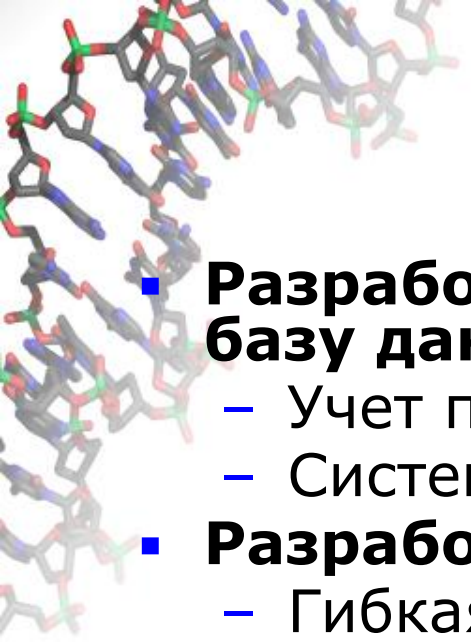
Membrane Proteins
- Sec and Translocase Proteins (5)
- ATP Binding Cassette (ABC) Tran ... (1)



Аннотированные базы данных (АБД)

- В настоящее время существует около 15 баз данных, которые содержат сведения о пространственных структурах РНК.
- **Не существует БД, которая обеспечивает аннотацию «неканонических» структур и возможность поиска по ним.**
- АБД пространственных структур РНК используются, в основном, для гомологического моделирования, подбора knowledge-based параметров и тестирования программ предсказания и сравнения структур РНК.

По нашему мнению, их потенциал для собственно биологических исследований используется недостаточно.

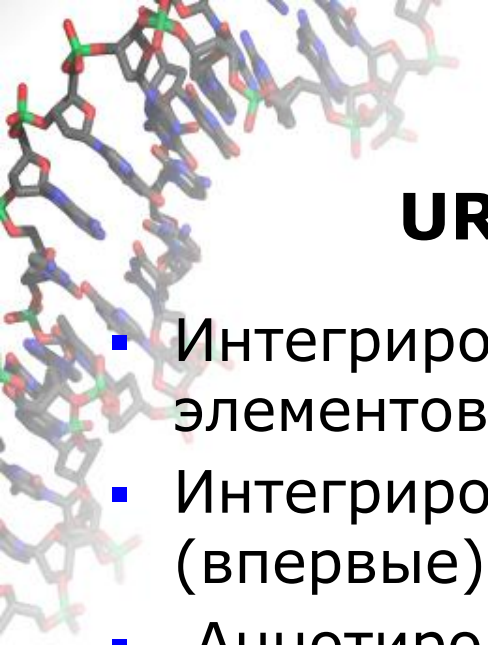


Задачи

- **Разработать универсальную аннотированную базу данных пространственных структур РНК**
  - Учет псевдоузловых структур;
  - Систематизация третичных мотивов.
- **Разработать веб-интерфейс базы данных**
  - Гибкая и понятная система поиска;
  - Поиск как полных структур, так и структурных элементов;
  - Сбор статистики.

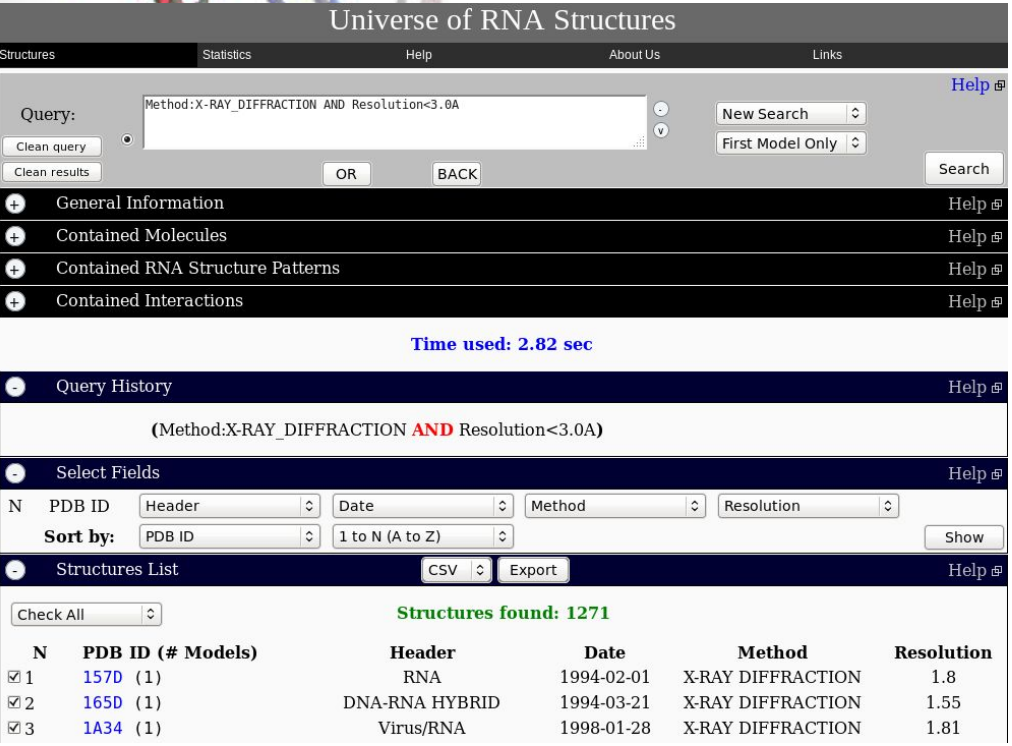
Исследовать структуры РНК, используя возможности базы

- Анализ свойств спиралей (длины, нукл. состав, роль в образовании псевдоузлов)
- **…**



URS – Universe of RNA Structures

- Интегрирована оригинальная классификация элементов вторичной структуры РНК;
- Интегрирована классификация псевдоузлов (впервые);
- Аннотированы элементы структур РНК:
  - Спаривания оснований и РНК-белковые контакты;
  - Нити, петли, крылья спиралей, спирали, линки;
  - Вторичные структуры, псевдоузлы.



URS – Universe of RNA Structures

Веб-интерфейс

- Анализ отдельной структуры;
  - Анализ отдельного структурного элемента;
  - Создание выборки структур (запрос в дизъюнктивной нормальной форме);
  - Сбор статистики структурных элементов из данной выборки.
- Доступен по адресу:

http://server3.lpm.org.ru/urs/

Дальнейшее развитие

- Неизбыточный список структур РНК
- Привлечение методов машинного обучения (в частности – для предсказания вторичной структуры)
- Использование базы данных для анализа структурных мотивов (сейчас – роль коротких стемов в псевдоузловых структурах)
- Улучшение аннотаций и веб-интерфейса

- Баулин Е. Ф., Ройтберг М. А., Астахова Т. В. Классификация элементов вторичной структуры РНК // Математическая биология и биоинформатика. 2012. Т. 7. № 2. С. 567-571;
- Баулин Е. Ф., Ройтберг М. А. Стемовые мультиплеты: новый подход к описанию третичных мотивов РНК // Математическая биология и биоинформатика. 2015. Т. 10. № 1. С. 54-59. doi: 10.17537/2015.10.54;
- Baulin E., Yacovlev V., Khachko D., Spirin S., Roytberg M. URS DataBase: Universe of RNA Structures and their motifs// Database **(на 2-м этапе рецензирования)**

Лаборатория прикладной математики

№№	ФИО, степень	Должность	Возраст
1	Ройтберг М.А., д.ф.-м.н.	Зав. лаб.	
2	Фурлетова Е.И., к.ф.-м.н.	Научный сотрудник	до 35 лет
3	Яковлев В.В., к.ф.-м.н.	Научный сотрудник	до 35 лет
4	Астахова Т.В.	Научный сотрудник	
5	Ройтберг А.М.	М.н.с.	до 35 лет
6	Карпов А.В.	Вед.програм.	
7	Хачко Д.В.	Инженер	до 35 лет
8	Максимова З.И.	Инженер	
9	Карпова Н.Ф.	Инженер	
10	Акименко А.В.	Математ.	
11	Баулин Е.Ф.	асп., ИМПБ РАН	до 35 лет
12	Андреев Р.В.	асп., ИПМ РАН	до 35 лет
13	Поверенная И.А.	асп. ФББ МГУ	до 35 лет
14	Горев Д.И.	магистр., ФИВТ МФТИ	до 35 лет

Спасибо за внимание !