В.А. Любецкий (lyubetsk@iitp.ru)

кафедра

«Математической логики и теории алгоритмов»

мех-мата МГУ (http://lpcs.math.msu.su/rus/staff.htm)

лаборатория

«Математических методов и моделей в биоинформатике»

Института проблем передачи информации РАН

(http://lab6.iitp.ru/ru/pub/)

По адресу http://lab6.iitp.ru/ru/pres/

лежит список собственно математических и информатических задач Биоинформатики – доклад В.А. Любецкого «Математические и computer science проблемы биоинформатики» *GraphHPC-2017*, МГУ, 2 марта 2017. См. также http://lab6.iitp.ru/ru/pub/

(Некоторая субъективность присутствует.)

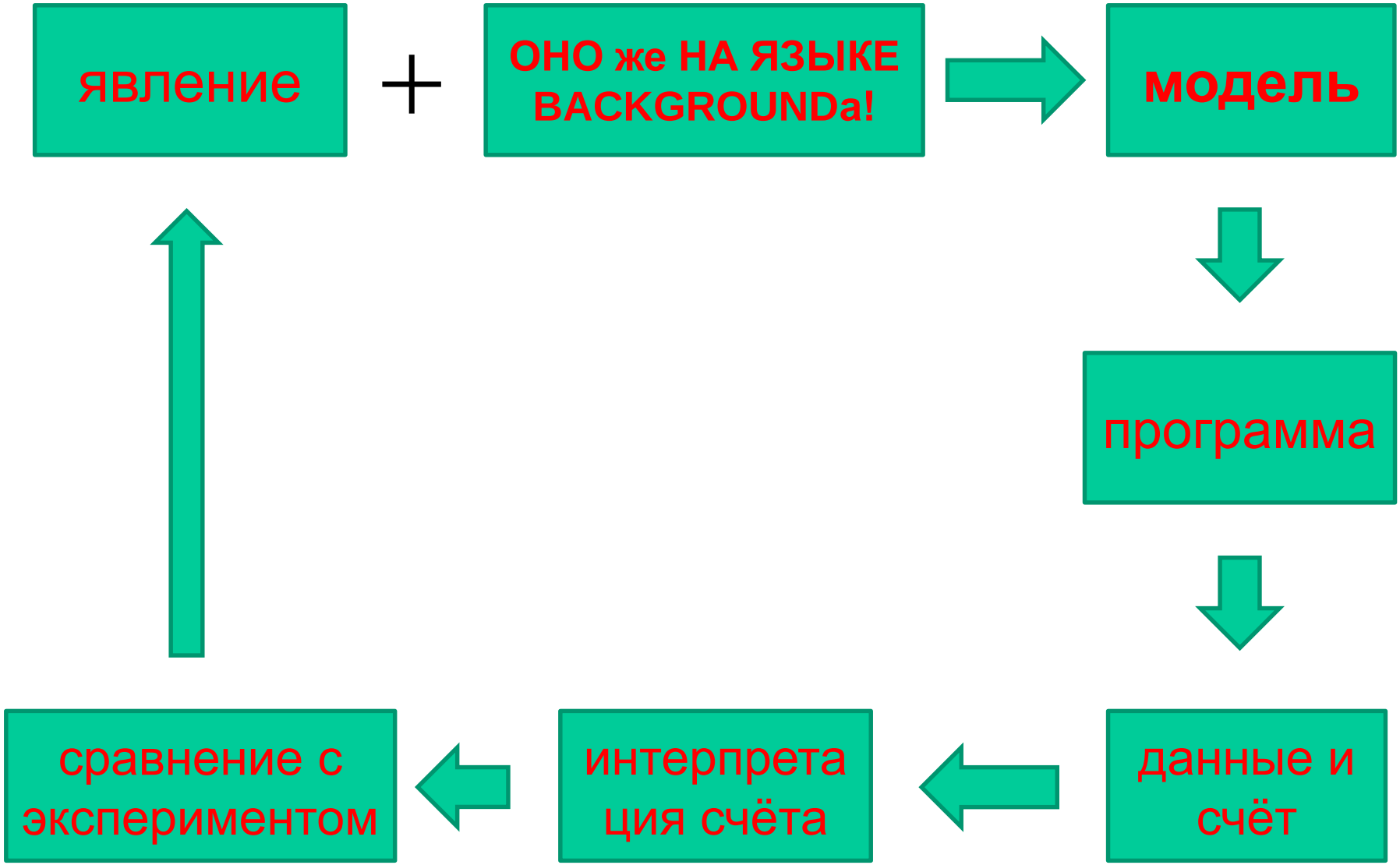
Но мне бы казалось правильным начать с более широкого Введения в этот Предмет: «Начало биоинформатики с точки зрения математики»

ВАКовская специальность 03.01.09:

«Математическая биология, Биоинформатика».

В названии указаны две части одного предмета, здесь хорошая аналогия с Математической физикой: «Математическая биология» – математические и информатические модели (биологических) явлений, Биоинформатика – счёт этих моделей (т.е.компьютерные программы и сам счёт на суперкомпьютере = распределённой системе).

Здесь предполагаются: (1) (биологические) ИСХОДНЫЕ ДАННЫЕ, т.е. пользование и/или развитие уже готовых и/или создание оригинальных Баз данных. Это – огромные Базы данных: длина генома x умножить на число y организмов (ныне живущих или когда-либо бывших, и их число непрерывно растёт)! До создания самой (3) МОДЕЛИ биологического явления нужен какой-то (2) BACKGROUND к используемым в модели основным понятиям. И этот Background должен быть математическим = строгим описанием! (4) ИНТЕРПРЕТАЦИЯ результатов счёта и их (5) СРАВНЕНИЕ с экспериментальными данными (которые содержатся в других базах данных и в статьях с экспериментами).


явление
+
ОНО же НА ЯЗЫКЕ BACKGROUNDa!
модель
программа
данные и счёт
интерпретация счёта
сравнение с экспериментом

РАБОЧИЙ ПЛАН ВОЗМОЖНОГО КУРСА/ПРАКТИКУМА:

ОБЩАЯ ЧАСТЬ: BACKGROUND.

1. Введение
2. Background

} Эти две части и составляют мой текущий доклад.

3. Базы данных: GenBank и Ensembl

(3а. Специфич. программирование: ANSI-C + MPI)

СПЕЦИАЛЬНАЯ ЧАСТЬ: МОДЕЛИ И АЛГОРИТМЫ ОСНОВНЫХ ЯВЛЕНИЙ: транскрипции, трансляции, эволюции.

4. Консервативные слова в геномах.
5. Выравнивание последовательностей.
6. Кластеризация белков.
7. http://lab6.iitp.ru/ru/pres/2017_graphhpc.pdf **и т.д.**

NCBI (Национальный центр биотехнологической информации http://www.ncbi.nlm.nih.gov/)
Содержит такие бесценные **ресурсы** как **GenBank** («аннотированная коллекция ***всех*** общедоступных последовательностей ДНК»), Reference Sequence (**RefSeq**) (неизбыточный хорошо аннотированный курируемый набор ***эталонных*** последовательностей геномов, транскриптов и белков), **PubMed** и PubMed Central (**PMC**) (база данных аннотаций статей биомедицинской тематики и архив их полных текстов), Sequence Read Archive (**SRA**) (хранилище «сырых» данных с секвенаторов «нового поколения») и мн. др.; а также **инструменты** для работы с ними, такие как Basic Local Alignment Search Tool (**BLAST**) (находит участки локального сходства биологических последовательностей) с многочисленными вариантами (которые студентам освоить) и Taxonomy Browser (навигатор по (довольно устаревшей) таксономии).

Ensembl (http://www.ensembl.org/) — хорошо структурированная БД, содержащая «избранные» полные геномы. Основной сайт посвящён **позвоночным** (~100 видов), но есть аналоги про других **животных** (~50), **растения** (~50), **грибы** (~600), **простейшие** (~200) и **бактерии** (> 40 000).

EuPathDB (http://eupathdb.org/) — БД, посвящённая геномам одноклеточных паразитов.

OrthoMCL DB (http://orthomcl.org/) — БД ортологичных групп. Содержит **1.5 млн белков** 150-ти организмов (представлены все формы жизни, включая вирусы), объединённые в ~**125 000 ортогрупп**.

UniProt (http://www.uniprot.org/) — БД белков и информации об их функциях.

Xfam (http://xfam.org/) — БД консервативных доменов белков (Pfam) и РНК (Rfam), повторов ДНК (Dfam), филогении белковых семейств (TreeFam), взаимодействия белков (iPfam) и др.

Genomicus (http://genomicus.biologie.ens.fr/) — инструмент исследования синтении у позвоночных.

Один из разделов Базы данных GenBank

(с него полезно начинать):

http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Undef&id=2759&lvl=3&keep=1&srchmode=1&unlock)

Теперь перейдём в разделу BACKGROUND,

который состоит из

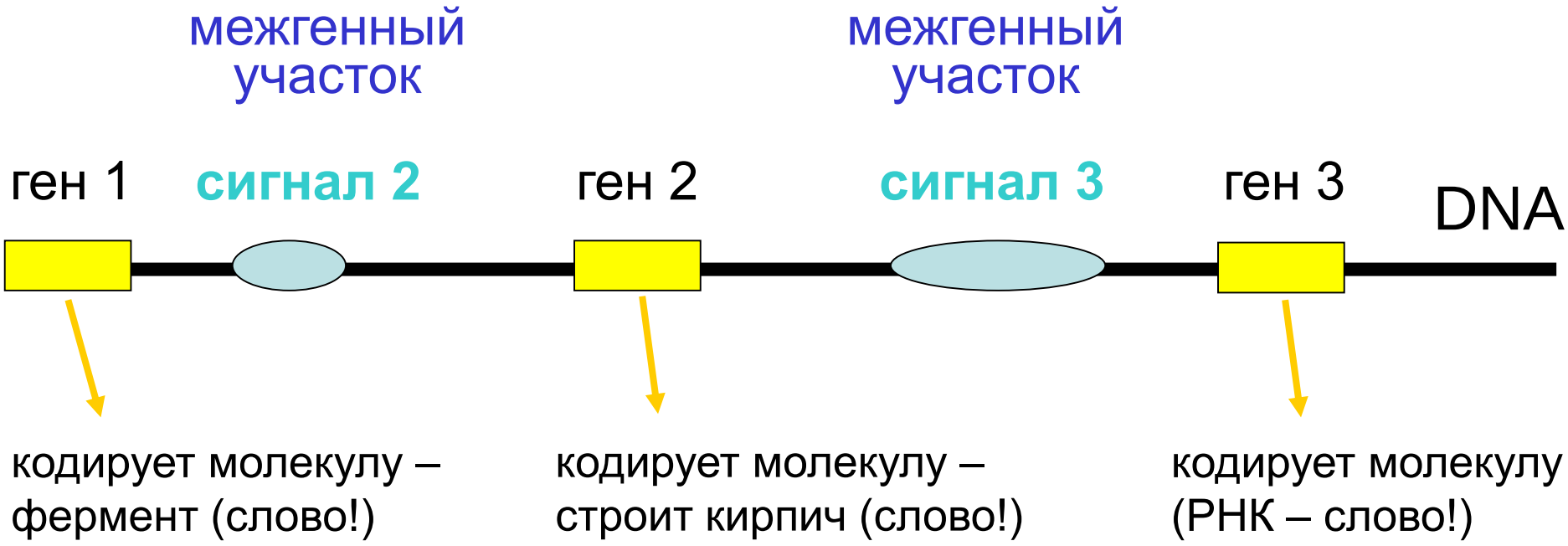
ГЕНОМИКИ и ФИЛОГЕНЕТИКИ.

1. **математический (= т.е. строгий)**

Background: геномика

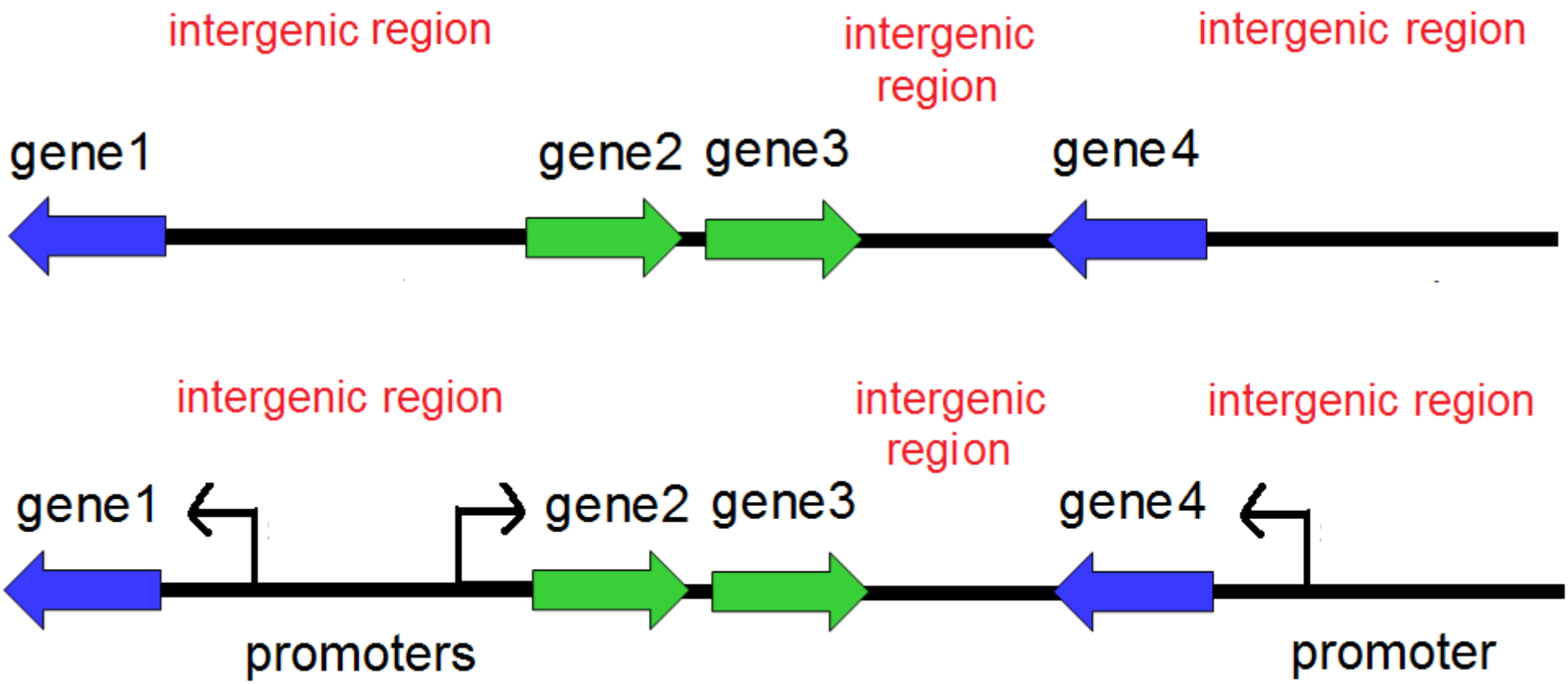
(рассказываю, в основном, о бактериях)

ДНК – последовательность в 4-буквенном алфавите {**A**,**C**,**T**,**G**} с характерной длиной 3 миллиона – 6 миллиардов (бывает и длина 17 тыс букв). Каждая **буква** называется **нуклеотид**: TTGACATGGCTATATAAGTCATGTTATACT - здесь 30 позиций



«Суть жизни»: ген (=слово) кодирует молекулу (=слово) и считывается или не считывается по сигналу (слову или системе слов) из межгенного участка

На самом деле: гены и сигналы – векторы!



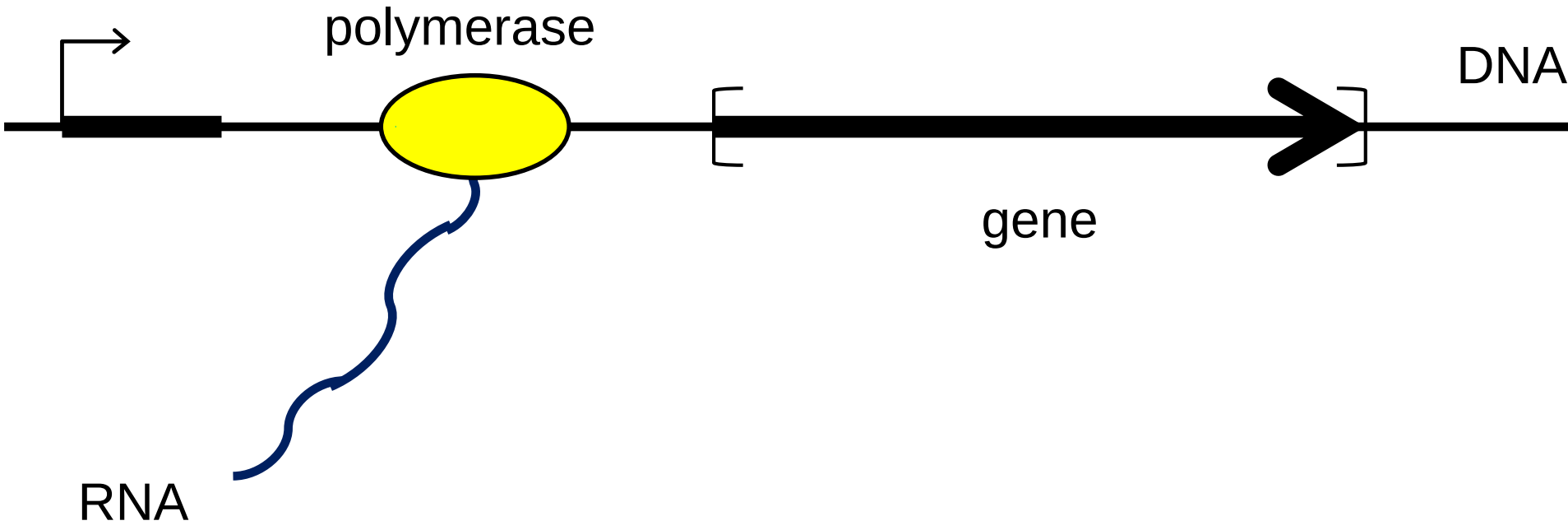
Пример **СИГНАЛА** (=участка с определённым буквенным составом и с направлением), здесь «промотора»:

human **CAAACCCCAAAGACA** →

frog **ACRTTATA** (R=A или G) →

bacteria **TTGACA -17..18- TATAAT -4..7- R** →

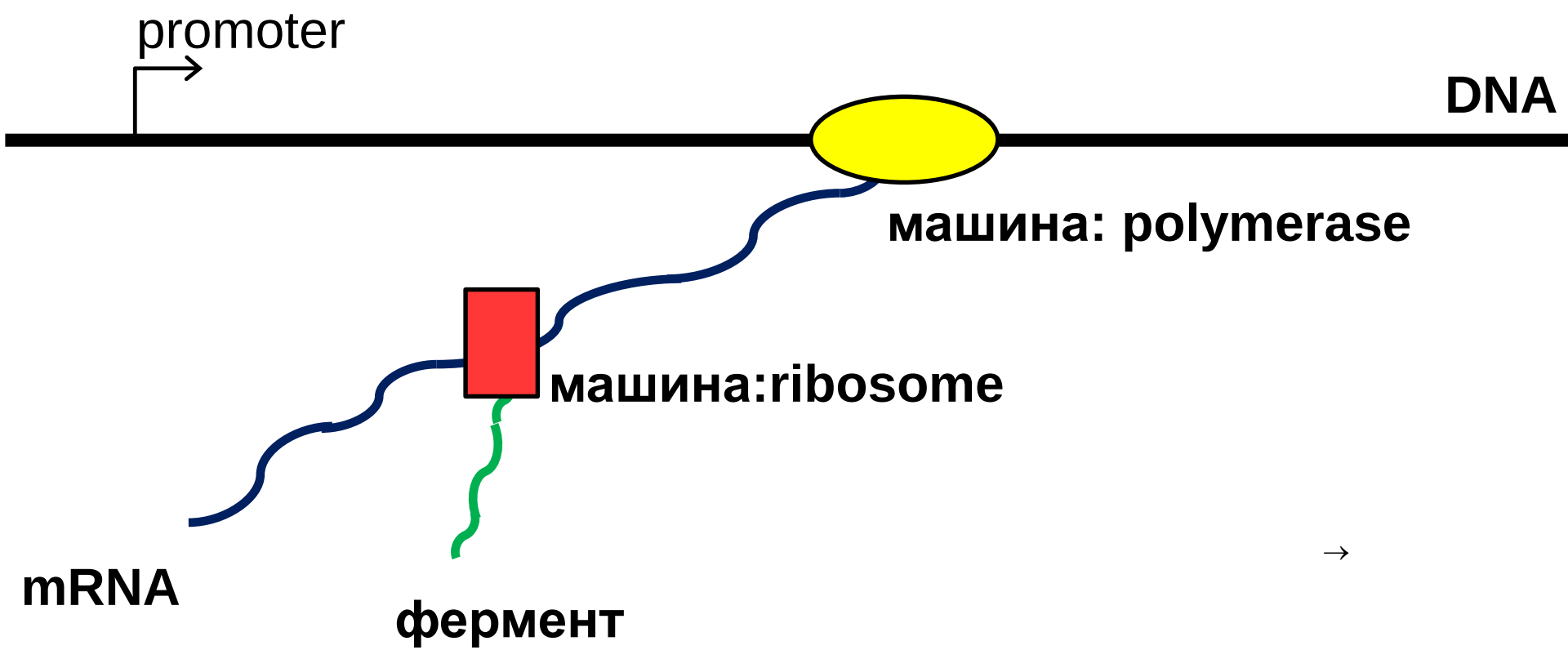
Что такое «чтение гена»? (= его «транскрипция»)



1й основной процесс в «текущей жизни клетки»: **связывание** с промотором молекулы, называемой РНК-**полимеразой** и её **движение** по направлению промотора. Если полимераза **проходит** любой **сонаправленный** с нею ген, она его **читает** («**копирует его в РНК**») по правилу «буква в букву».

ДНК ⇒ **РНК – слово в 4-х буквенном алфавите {A, C, U, G}**

ДНК – длинная последовательность, а РНК короткая!
(как память у комп: оперативная или на жёстком диске).

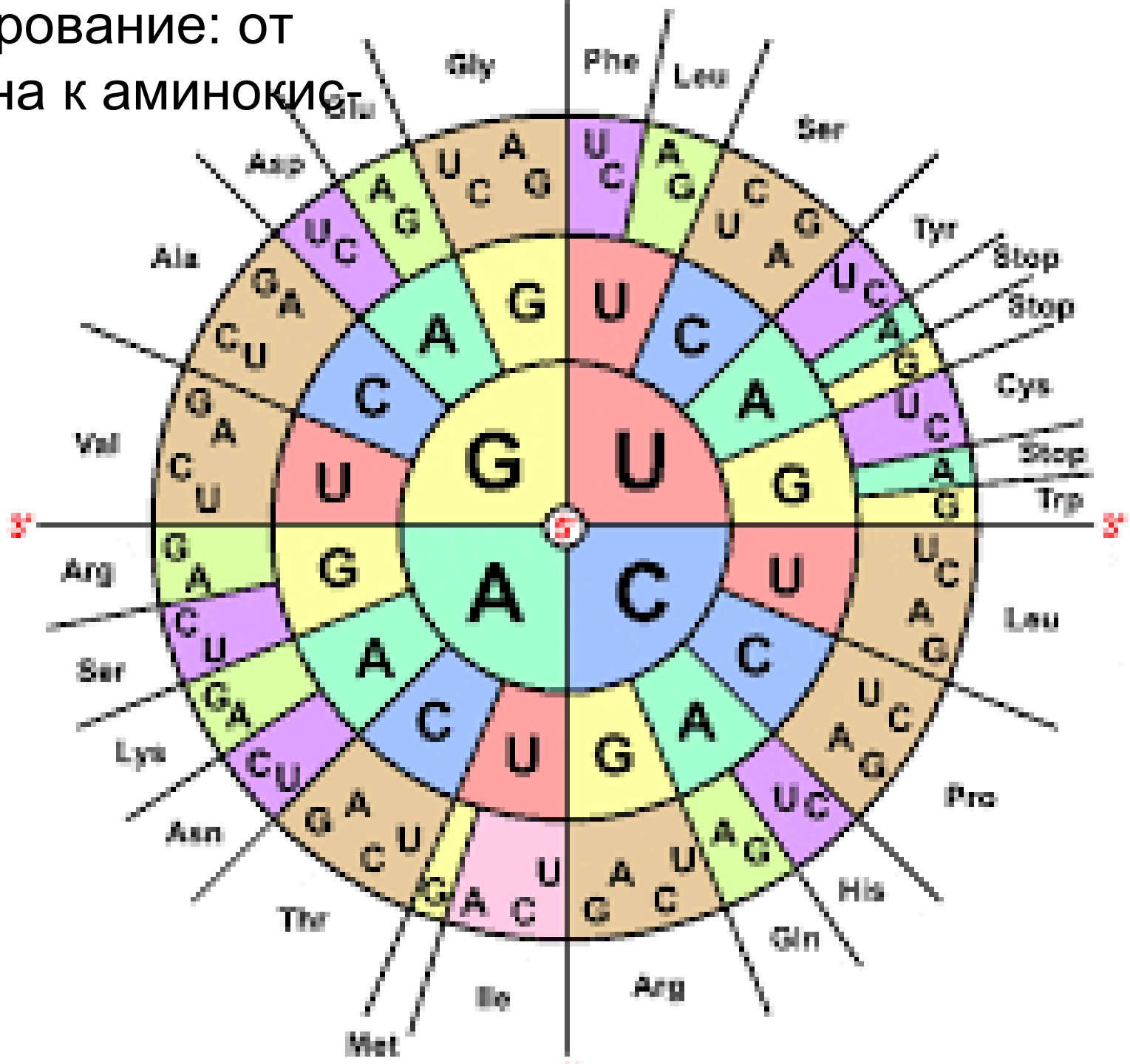


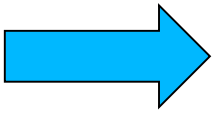
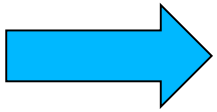
ФЕРМЕНТ – совсем короткая последовательность в 20ти бук-венном алфавите. Итак, имеется **два** основных процесса в «текущей жизни клетки»: **транскрипция** и затем **2й** процесс: **трансляция**. Фиксирована таблица кодировки: 3 → 1

Итак, рибосома перекодирует слово в 4х буквенном алфавите {A,C,U,G} нуклеотидов в слово в 20ти буквенном алфавите
(его буквы называются «аминокислотами»)
в соответствии с
кодовой ТАБЛИЦЕЙ, единой для всего живого
(почти).

А именно, соседние тройки нуклеотидов («кодоны»)
заменяются на одну аминокислоту по этой Таблице.

Кодирование: от кодона к аминокислоте:



В итоге СУТЬ ЖИЗНИ: ДНК  РНК  белок!

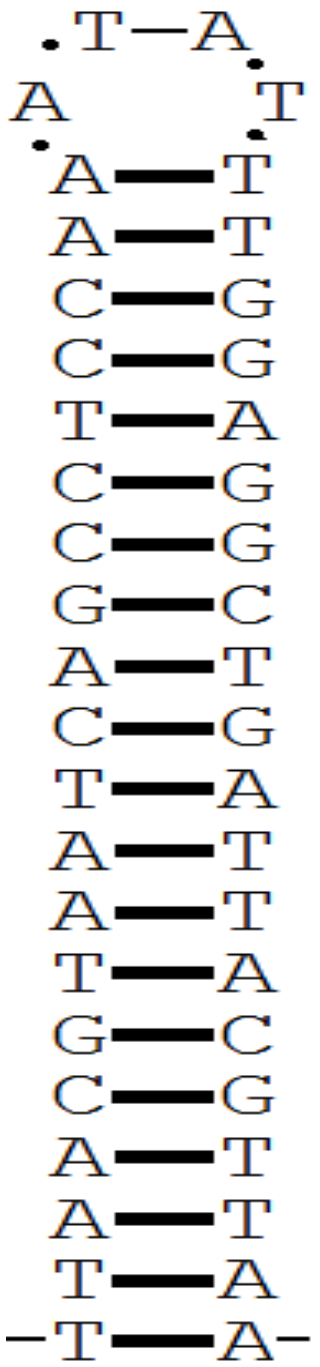
ЗАЧЕМ: как используются эти два перекодирования? В чём здесь «суть»?

ДНК находится внутри мешочка («цитоплазмы»), в котором идут химические реакции. Они идут цепочками друг за другом, но **не идут без ферментов** (см. слайд 31), а **ферменты кодируются генами**.

Итак, **ДНК с помощью ферментов управляет химическими реакциями в клетке, т.е. всей химической жизнью в клетке (и даже между клетками)**.

в РНК скрыт ещё один, вполне математический аспект – её вторичная структура:

Снизу показан участок РНК: жёлтая часть называется «**левым плечом**», голубая часть – «**правым плечом**»; между ними, – «**петля**». Справа показано, как такой участок сворачивается в «**СПИРАЛЬ**», образуя геометрию.

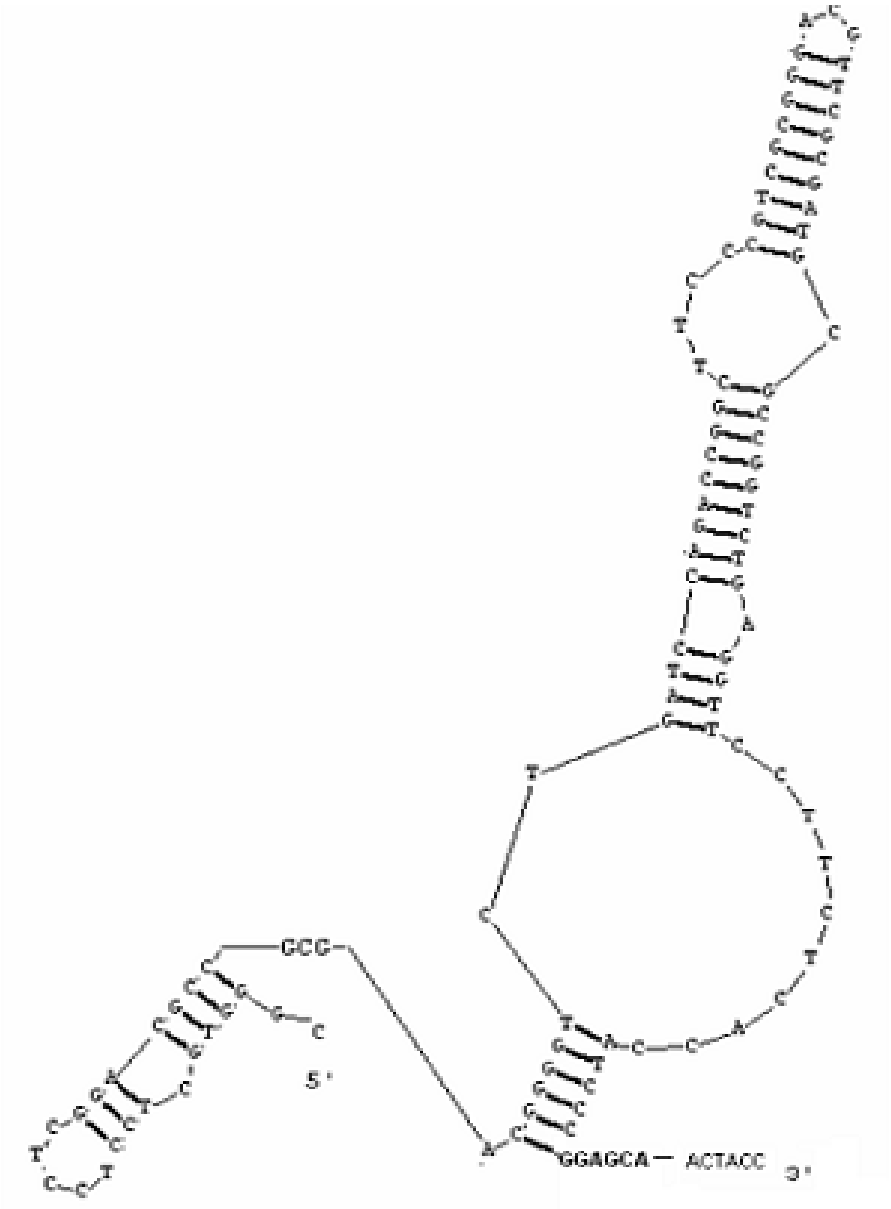


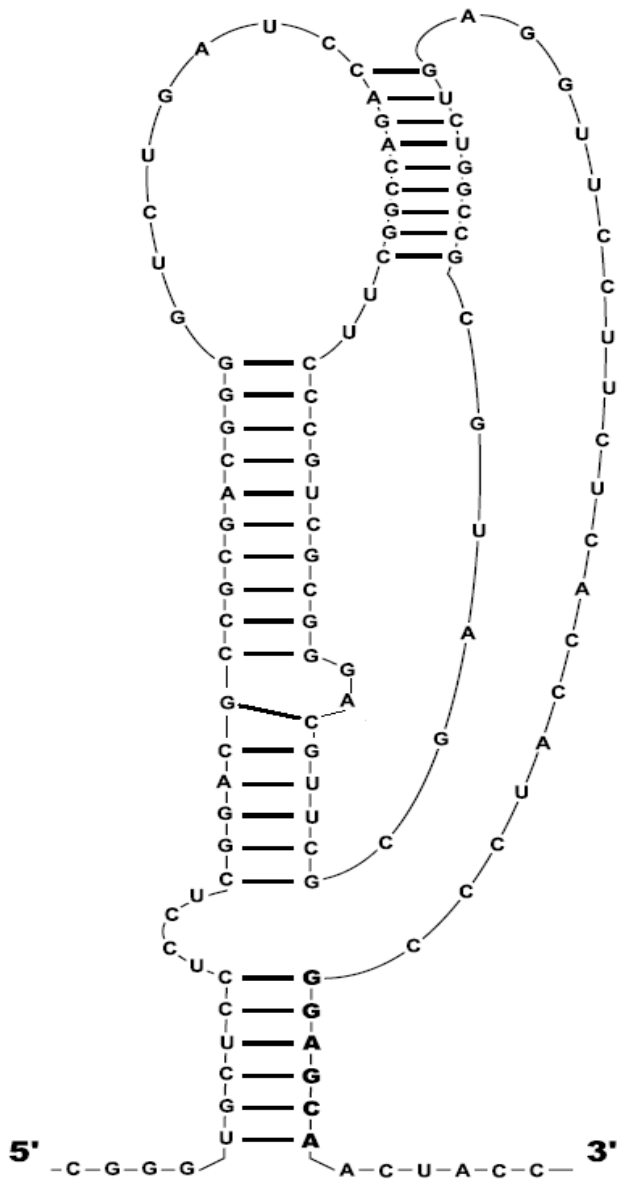
Сотни, тысячи спиралей образуют **ВТОРИЧНУЮ СТРУКТУРУ РНК**.

TTAACGTAATCAGCCTCCAAATATTTGGAGGCTGATTACGTTAA

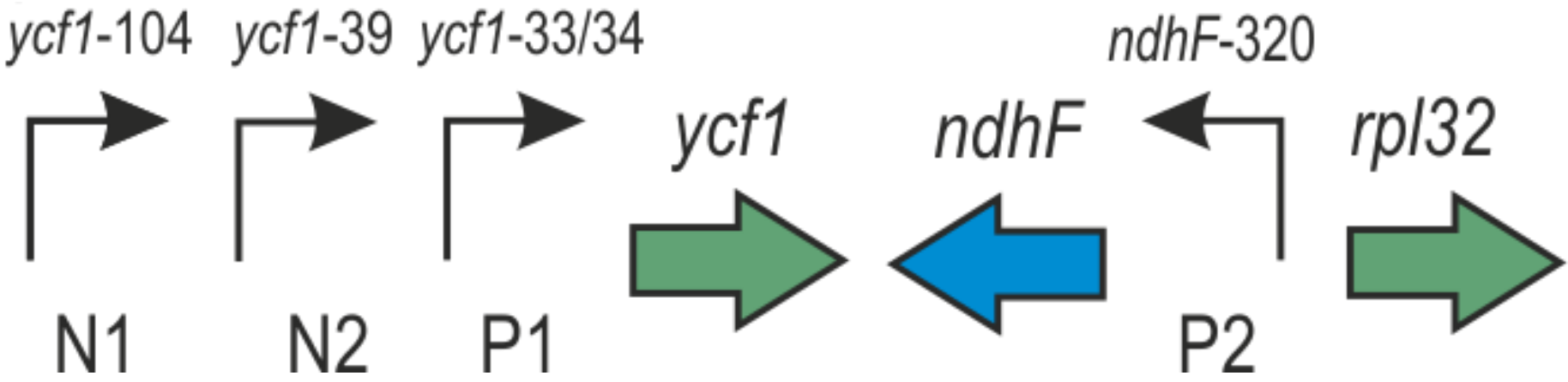
Итак, РНК – короткая последовательность (=слово) вместе с богатой вторичной структурой на ней.

Примеры простых **вторичных структур**





Пример: 3 гена и 4 промотора

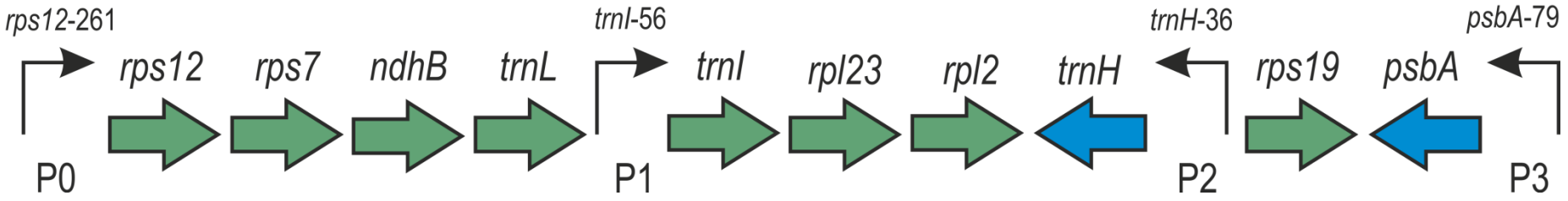


Взаиморасположение генов и промоторов (и других сигналов) может быть очень разнообразным.

В этом одна из трудностей математического изучения транскрипции

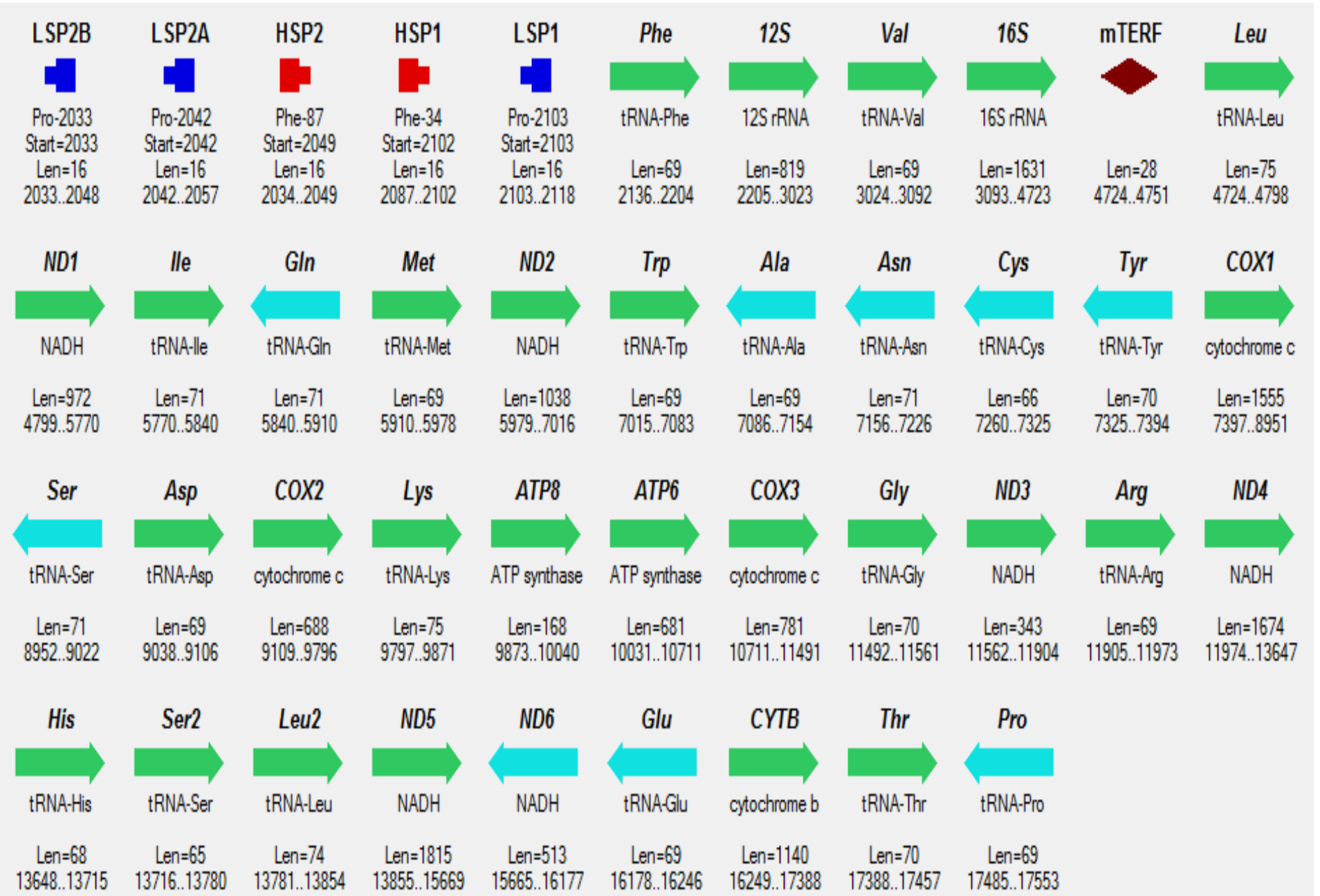
(геометрия бывает очень разной)

Еще пример: 10 генов и 4 промотора



Митохондриальный геном человека:

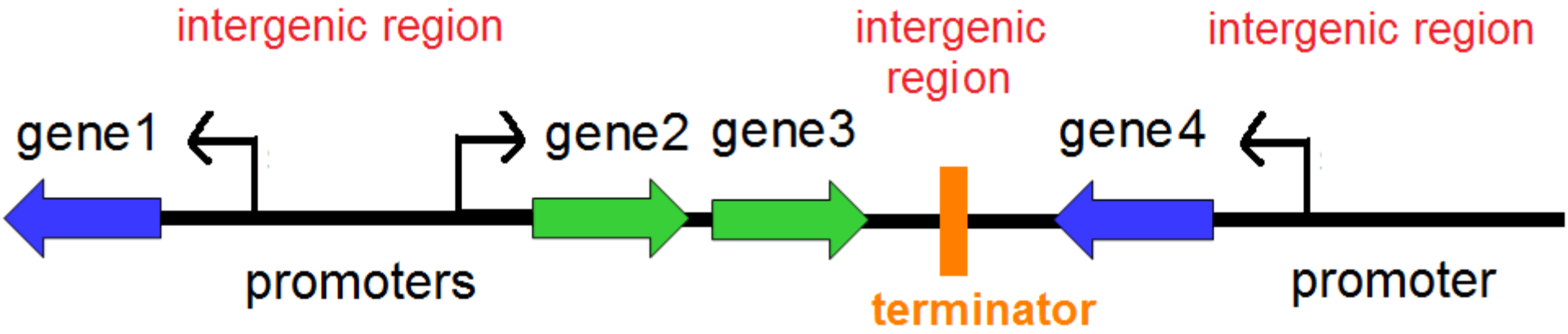
Митохондриальный геном лягушки:



Митохондриальный геном крысы:

Ещё одно действующее лицо: **терминатор**,

он **скидывает** определённую долю ***p*** полимераз, идущих с одной стороны, и определённую долю ***q*** полимераз, идущих с другой стороны (как два автоинспектора на разных сторонах дороги).



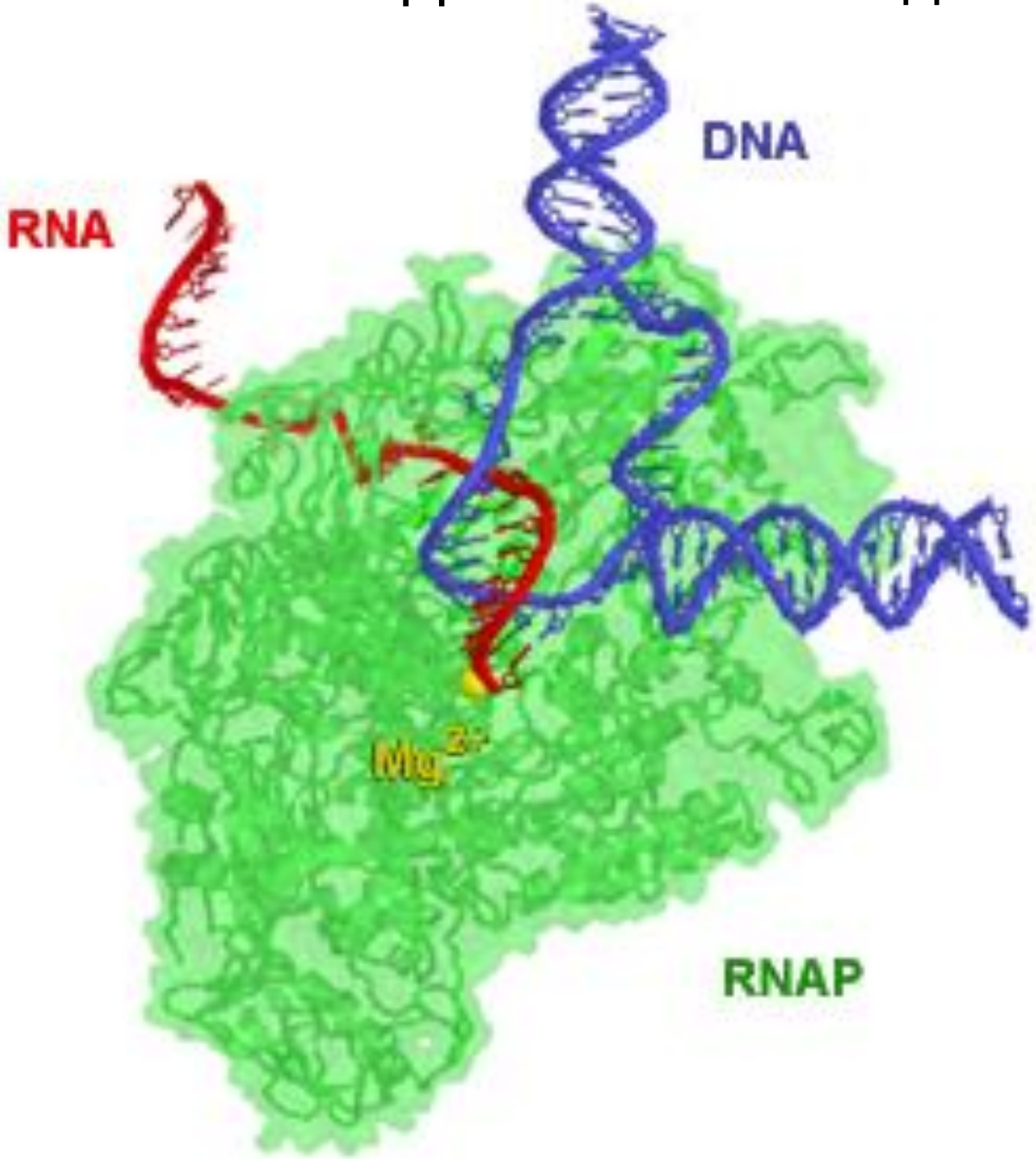
Различие в числах ***p*** и ***q*** называется степенью поляризации терминатора. Если одно из них равно 1,то он 100%-поляризованный в соответствующем направлении

Результат счёта на нашем суперкомпьютере всей этой сложной динамики, например, для лягушки лежит здесь:

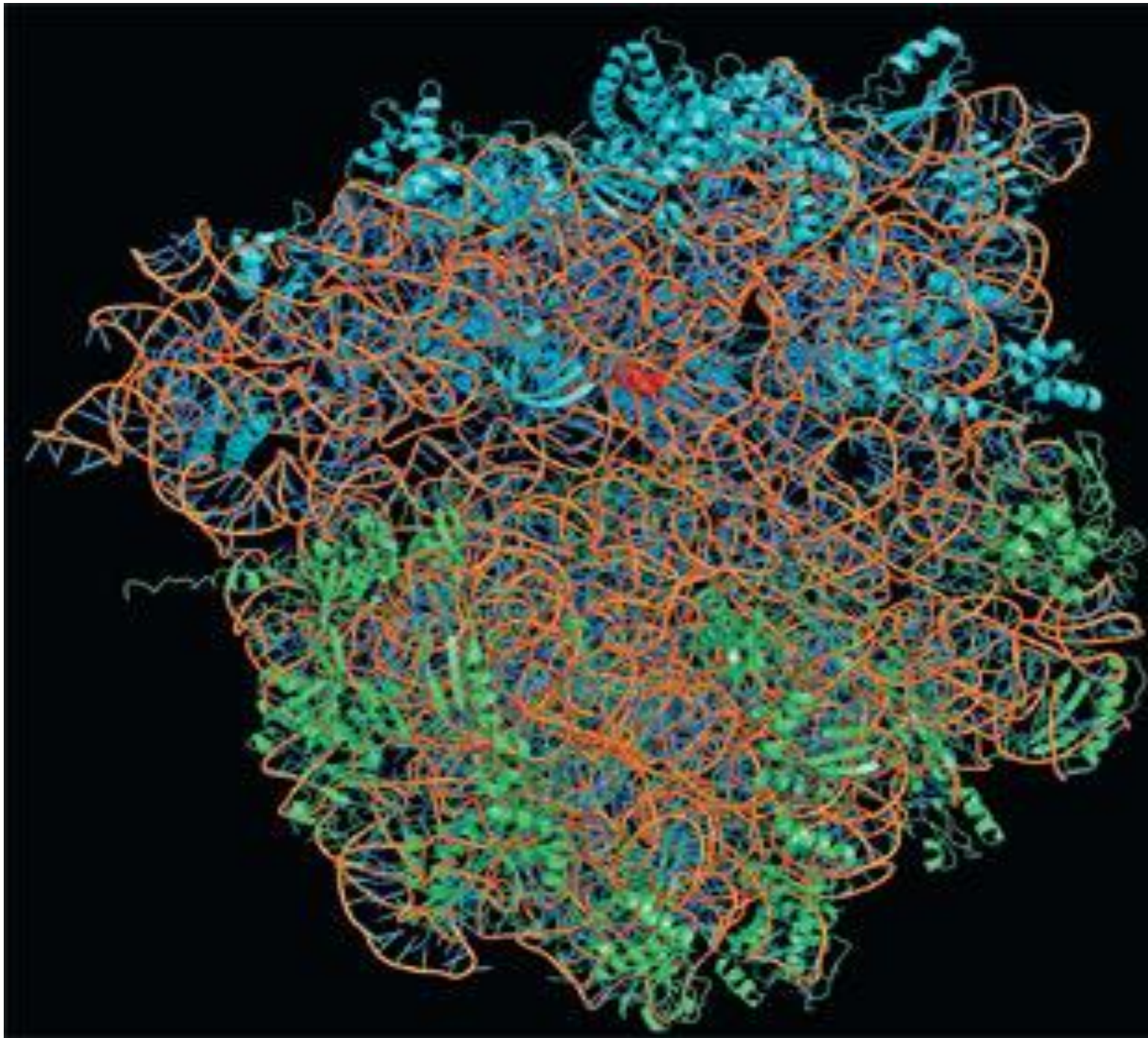
http://lab6.iitp.ru/docs/rivals/xlm.rar

Было бы очень важно теоретически изучить эту динамику: возможные режимы функционирования, «точки» переключения режимов и т.д.

РНК-полимераза ползёт по ДНК из неё выходит РНК:

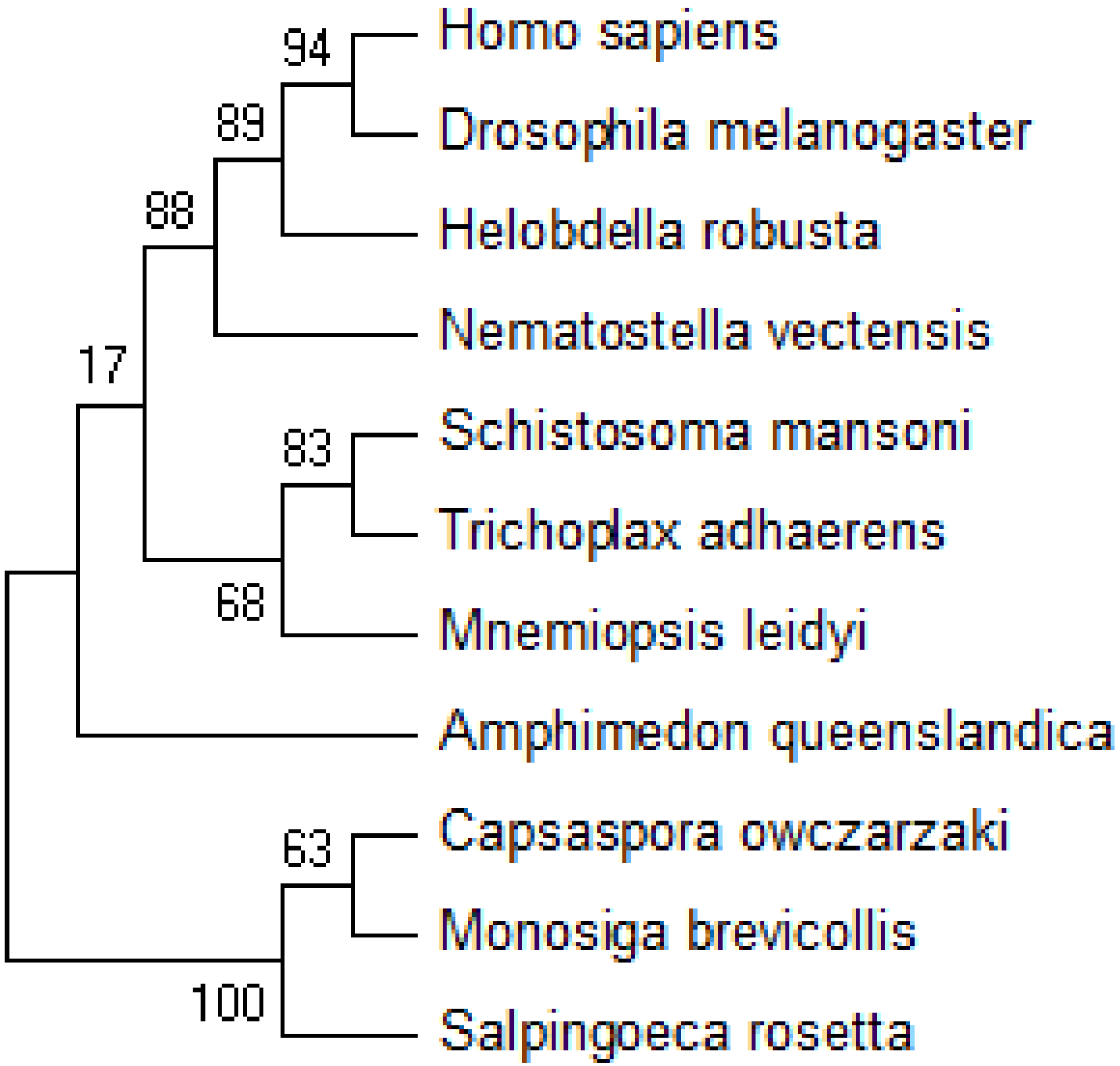


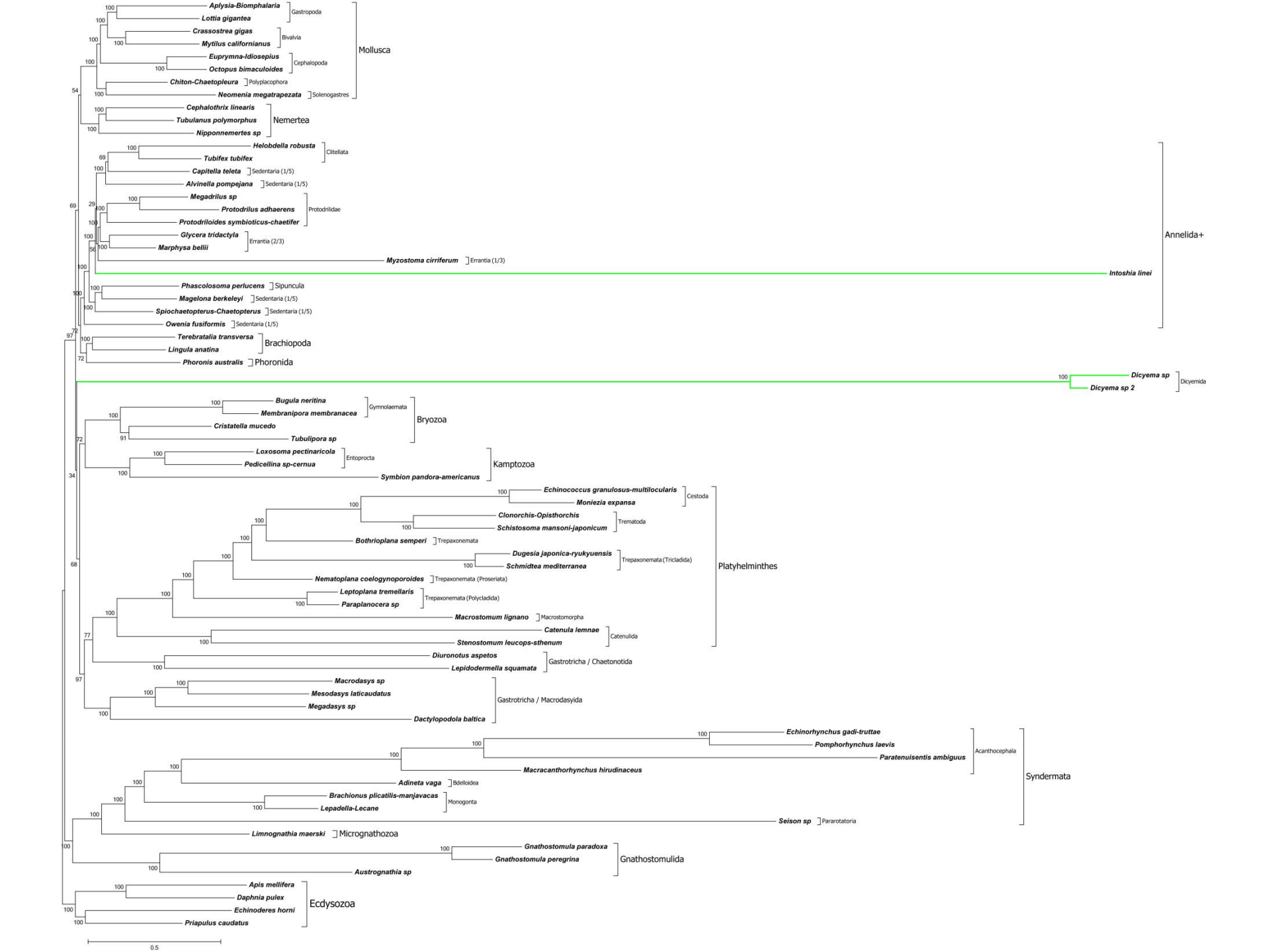
Рибосома ползёт по РНК и из неё выходит фермент:



Хим реакции в клетке: взаимодействие молекул (=слов!!)

2. Background: филогенетика


94
89
88
17
83
68
63
100
Homo sapiens
Drosophila melanogaster
Helobdella robusta
Nematostella vectensis
Schistosoma mansoni
Trichoplax adhaerens
Mnemiopsis leidyi
Amphimedon queenslandica
Capsaspora owczarzaki
Monosiga brevicollis
Salpingoeca rosetta


Aplysia-Biomphalaria
Lottia gigantea
Gastropoda
Crassostrea gigas
Mytilus californianus
Bivalvia
Euprymna-Idiosepius
Octopus bimaculoides
Cephalopoda
Chiton-Chaetopleura
Polyplacophora
Neomenia megatrapezata
Solenogastres
Mollusca
Cephalothrix linearis
Tubulanus polymorphus
Nipponnemertes sp
Nemertea
Helobdella robusta
Tubifex tubifex
Clitellata
Capitella teleta
Sedentaria (1/5)
Alvinella pompejana
Sedentaria (1/5)
Megadrilus sp
Protodrilus adhaerens
Protodriloides symbioticus-chaetifer
Protodrilidae
Glycera tridactyla
Marphysa bellii
Errantia (2/3)
Myzostoma cirriferum
Errantia (1/3)
Intoshia linei
Annelida+
Phascolosoma perlucens
Sipuncula
Magelona berkeleyi
Sedentaria (1/5)
Spiochaetopterus-Chaetopterus
Sedentaria (1/5)
Owenia fusiformis
Sedentaria (1/5)
Terebratalia transversa
Lingula anatina
Brachiopoda
Phoronis australis
Phoronida
Dicyema sp
Dicyema sp 2
Dicyemida
Bugula neritina
Membranipora membranacea
Gymnolaemata
Cristatella mucedo
Tubulipora sp
Bryozoa
Loxosoma pectinaricola
Pedicellina sp-cernua
Entoprocta
Symbion pandora-americanus
Kamptozoa
Echinococcus granulosus-multilocularis
Moniezia expansa
Cestoda
Clonorchis-Opisthorchis
Schistosoma mansoni-japonicum
Trematoda
Bothrioplana semperi
Trepaxonemata
Dugesia japonica-ryukyuensis
Schmidtea mediterranea
Trepaxonemata (Tricladida)
Nematoplana coelogynoporoides
Trepaxonemata (Proseriata)
Leptoplana tremellaris
Paraplanocera sp
Trepaxonemata (Polycladida)
Macrostomum lignano
Macrostomorpha
Catenula lemnae
Stenostomum leucops-sthenum
Catenulida
Platyhelminthes
Diuronotus aspetos
Lepidodermella squamata
Gastrotricha / Chaetonotida
Macrodasys sp
Mesodasys laticaudatus
Megadasys sp
Dactylopodola baltica
Gastrotricha / Macrodasyida
Echinorhynchus gadi-truttae
Pomphorhynchus laevis
Paratenuisentis ambiguus
Macracanthorhynchus hirudinaceus
Acanthocephala
Adineta vaga
Bdelloidea
Brachionus plicatilis-manjavacas
Lepadella-Lecane
Monogonta
Seison sp
Pararotatoria
Syndermata
Limnognathia maerski
Micrognathozoa
Gnathostomula paradoxa
Gnathostomula peregrina
Austrognathia sp
Gnathostomulida
Apis mellifera
Daphnia pulex
Echinoderes horni
Priapulus caudatus
Ecdysozoa
0.5

Как дерево строится? Строго математически!

А именно, (1) из данного списка организмов выбирают по 100 белков; (2) их конкатенируют, получается, что каждому организму соответствует одна длинная последовательность; вычисляют расстояние *r*(*i*,*j*) между этими последовательностями (= организмами) *i* и *j* <получается матрица *r*(*i*,*j*)>; (3) ищут дерево *T* (с длинами рёбер или без них), для которого невязка расстояний по дереву и по матрице минимальная.

Белки должны быть соответствую-ми («ортологичными»).

Эволюция (=дерево, граф) всего Живого:

Спасибо за внимание