Мотивы в ДНК: идентификация и практические приложения в задачах регуляторной геномики

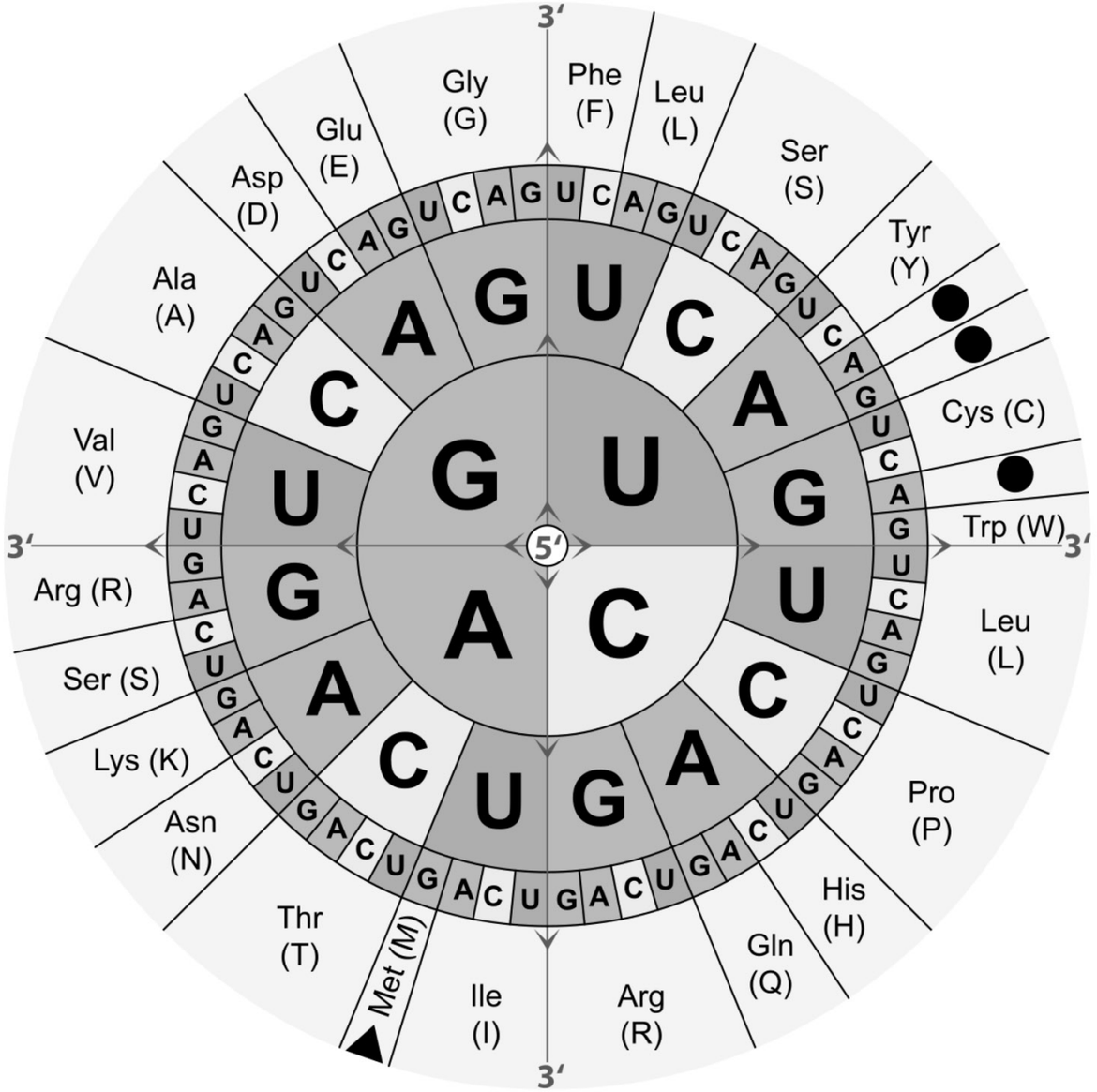
Ваня Кулаковский
ИМБ РАН, ИОГен РАН, Сколтех, autosome.ru

Robert W. Holley

Har Gobind Khorana

Marshall W. Nirenberg

Нобелевская премия (физиология и медицина) 1968 за "интерпретацию генетического кода и его функции в биосинтезе белка".


3′
Phe (F)
Leu (L)
Ser (S)
Tyr (Y)
Cys (C)
Trp (W)
Leu (L)
Pro (P)
His (H)
Gln (Q)
Arg (R)
Ile (I)
Met (M)
Thr (T)
Asn (N)
Lys (K)
Ser (S)
Arg (R)
Val (V)
Ala (A)
Asp (D)
Glu (E)
Gly (G)
5′
U C A G
G U C A
A G U C
C A G U

Генетический код
и его полезные свойства:

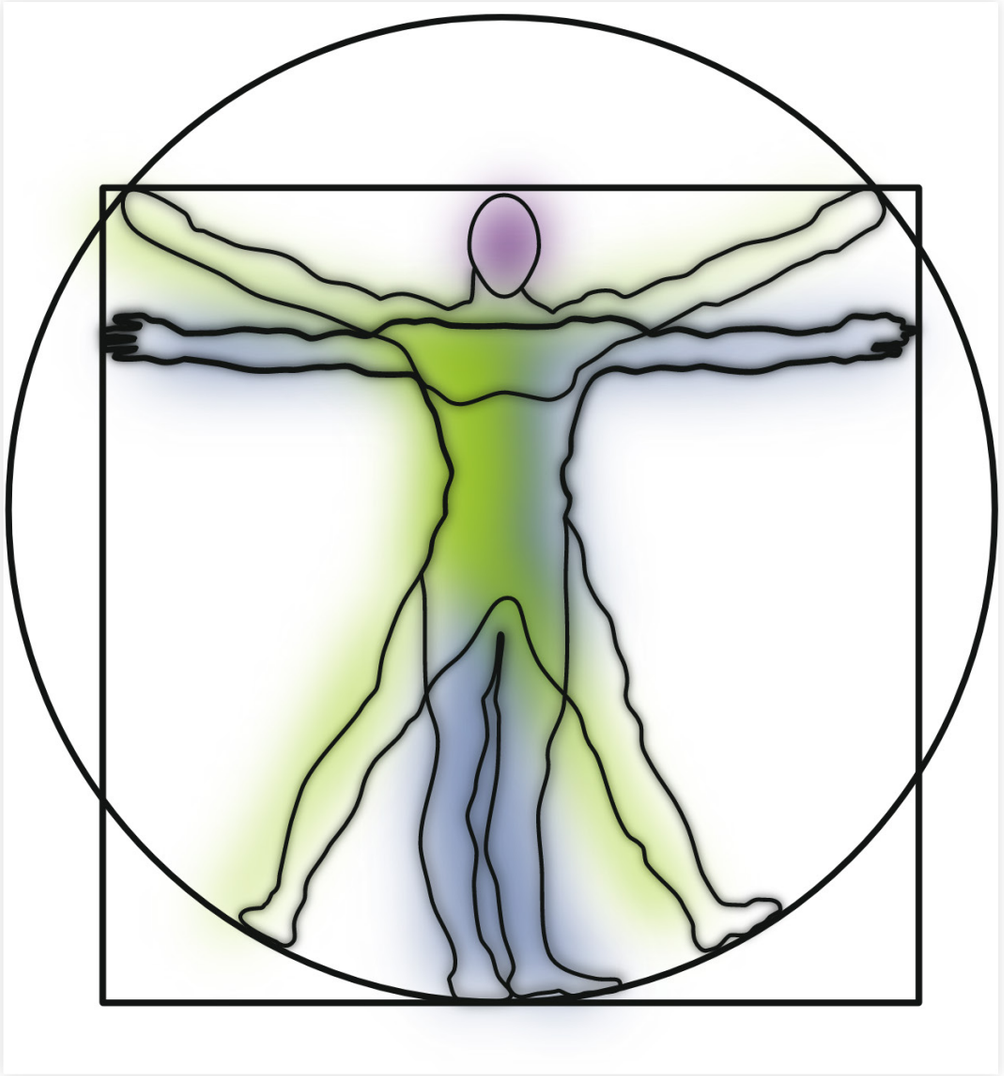
избыточность и вырожденность > синонимичные и несинонимичные замены > функциональная аннотация мутаций > молекулярные часы, оценки dN/dS > молекулярная эволюция

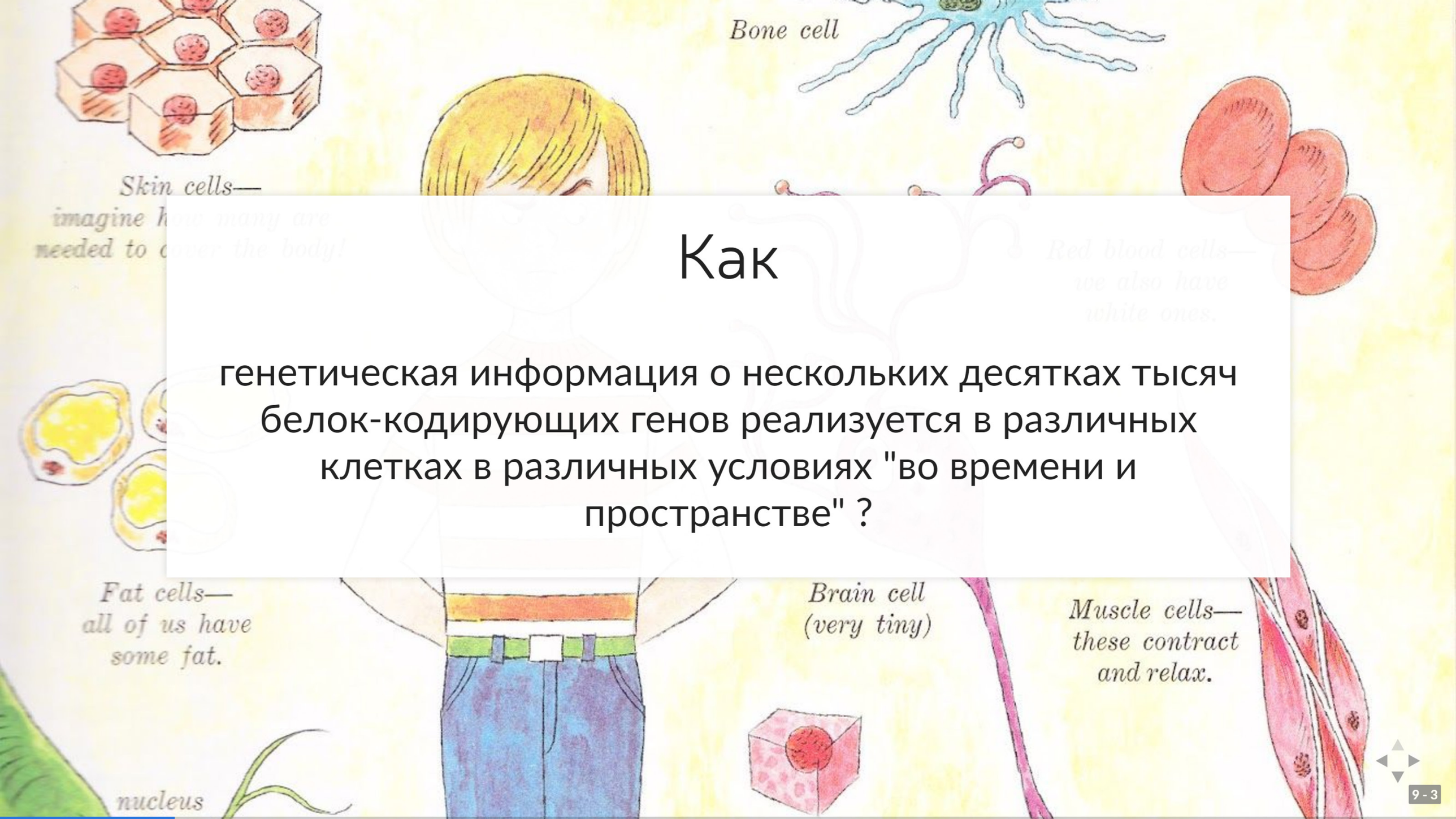
Геном человека

содержит "всего лишь" ~19000 белок-кодирующих генов
[HGNC, 2015 - 2017]

см. также Ezkurdia *et al.*, *Human Mol Gen*, 2014

What makes up a human? / Из чего же, из чего же сделаны наши мальчишки?





Как

генетическая информация о нескольких десятках тысяч белок-кодирующих генов реализуется в различных клетках в различных условиях "во времени и пространстве" ?

Вероятно, дело в *регуляции* экспрессии генов.

Существует ли *регуляторный* код?

Почему интересно изучать регуляцию у высших эукариот?

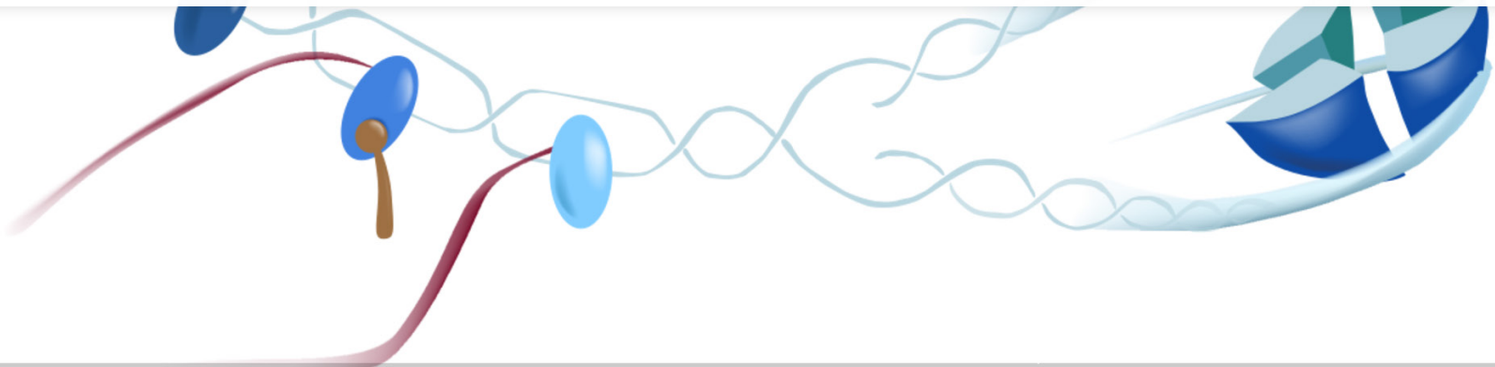
Процент экзонов от общей длины генома (Human Percent of Genome that is Exonic) - **2.8%**

Доля некодирующей ДНК в геноме кишечной палочки (Non-coding DNA in *E. coli*) - **12%**

[Harvard BioNumbers Database]

Уровни регуляции экспрессии генов

Упаковка ДНК > Гистоновый код > Метилирование ДНК > **Регуляция транскрипции факторами транскрипции** > *Регуляция сплайсинга > Регуляция трансляции > Контроль стабильности мРНК*



Regulated transcription controls the diversity, developmental pathways, and spatial organisation of the hundreds of cell types that make up a mammal.

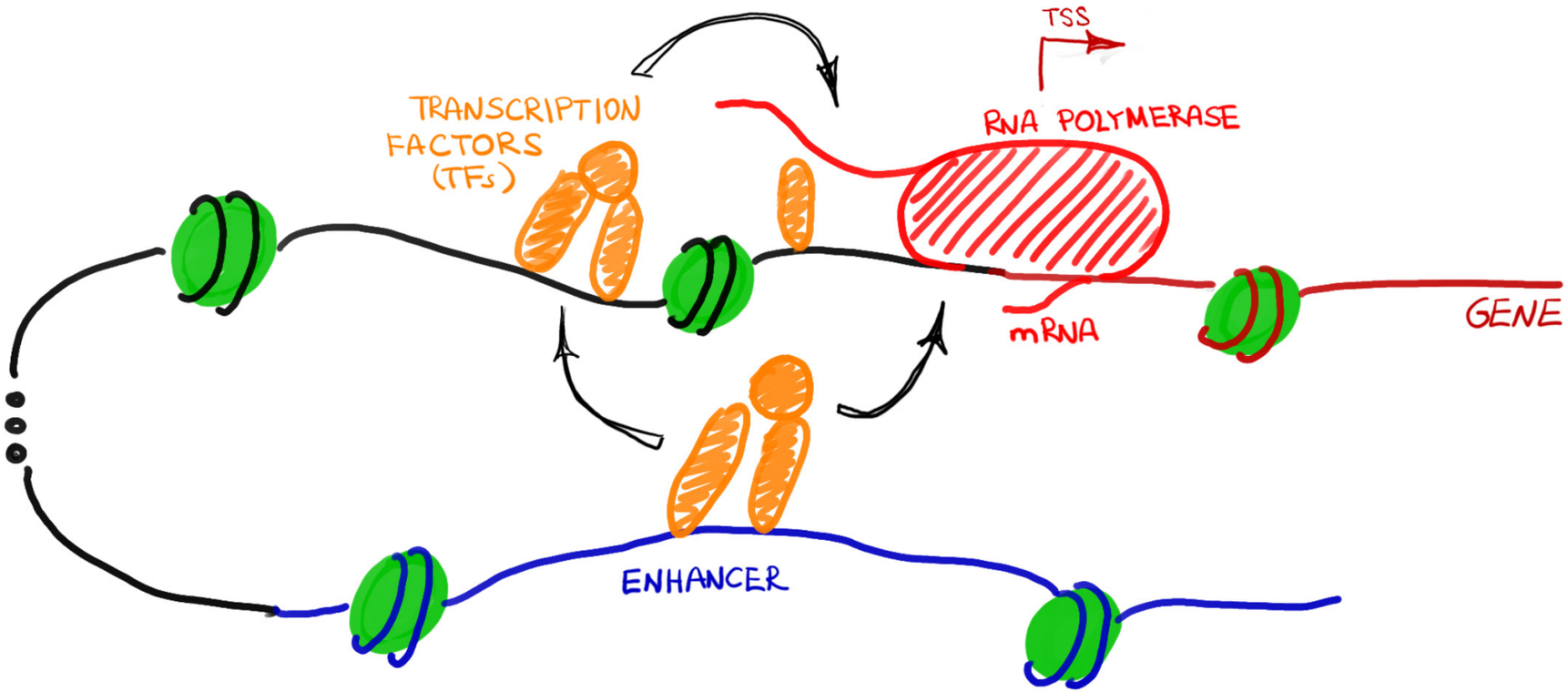
Проект FANTOM5, посвященный масштабному изучению инициации транскрипции у эукариот

A.R.R. Forrest *et al.*, *Nature*, 2014

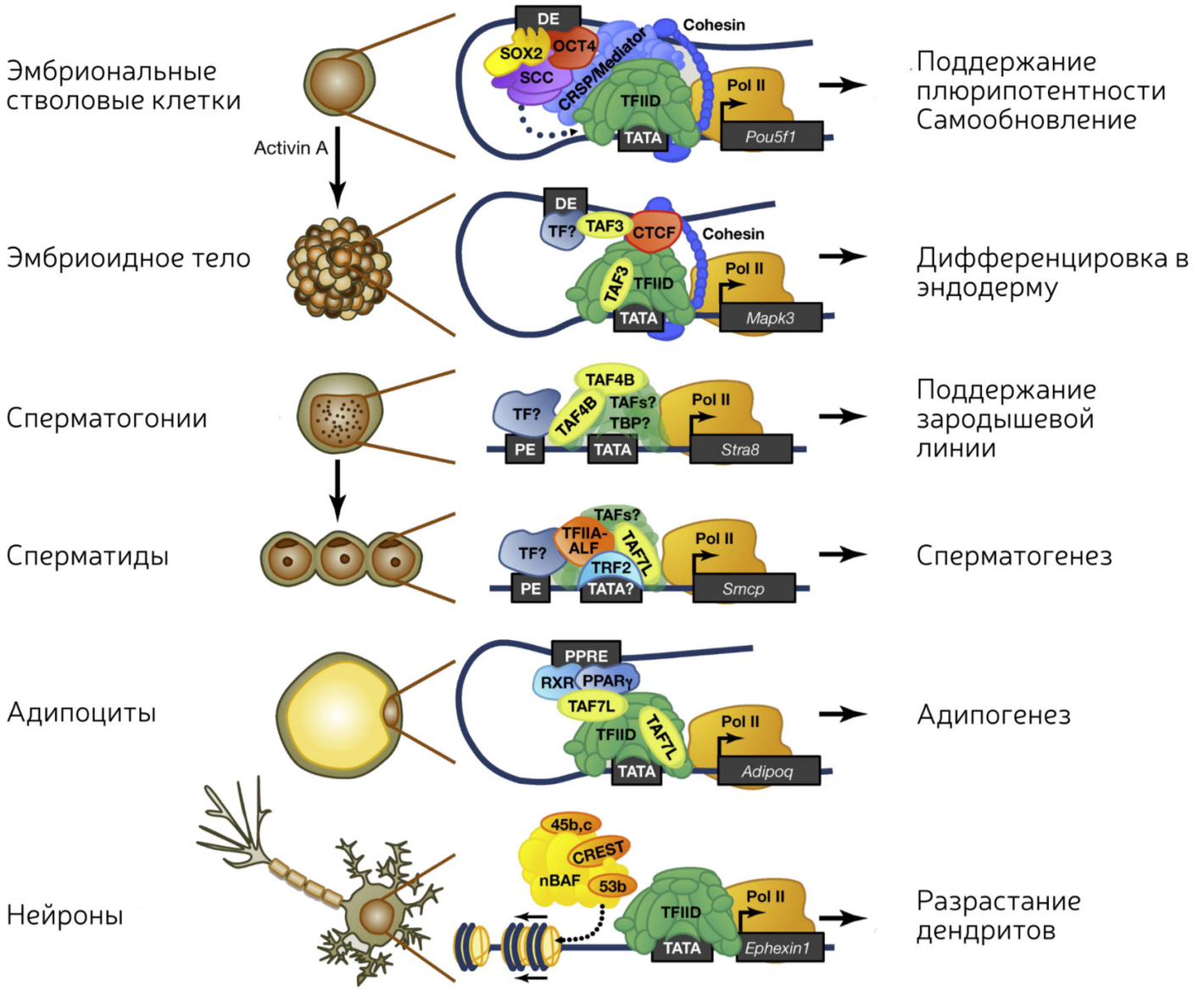
Transcriptional regulation in eukaryotes is *complex.*

Luke A. Gilbert *et al.*, *Cell*, 2013; Andrew J. Bonham *et al.*, *NAR*, 2013

Петлевая модель регуляции

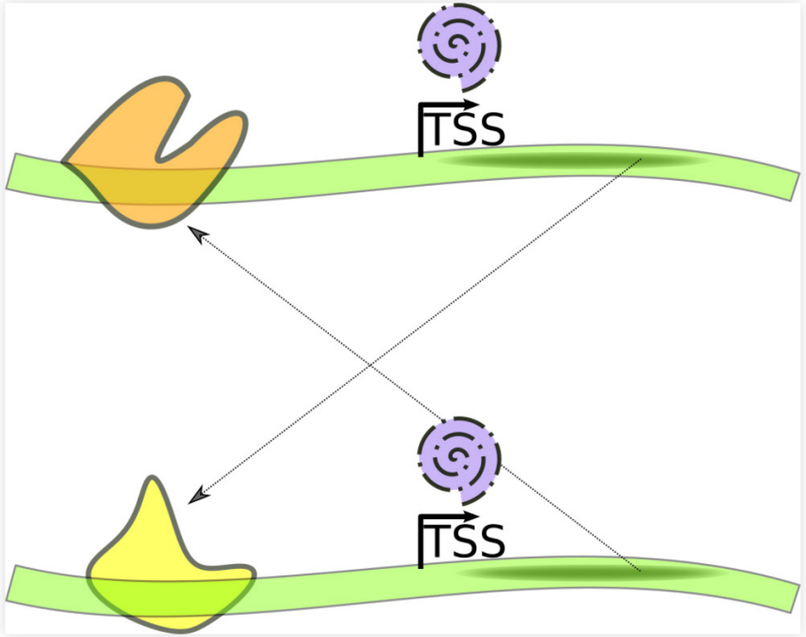


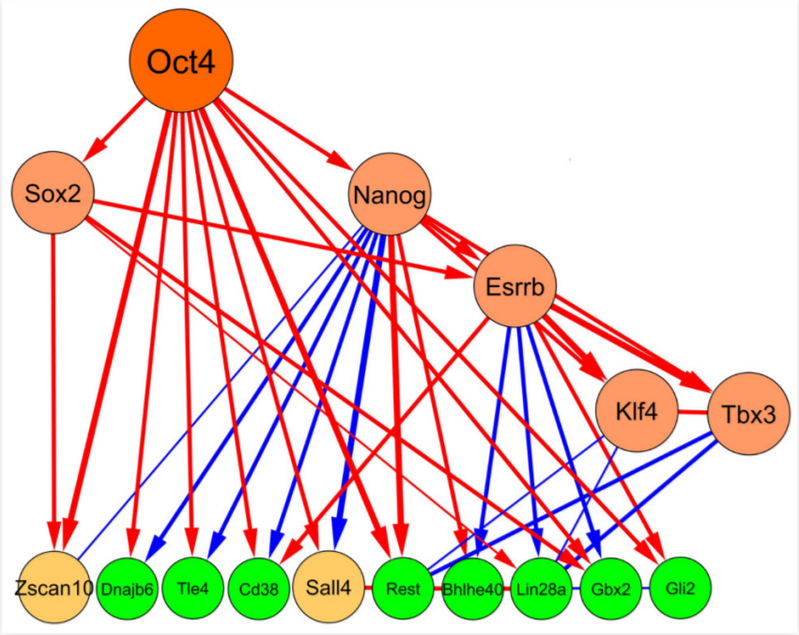
Факторы транскрипции контролируют и поддерживают "клеточную идентичность"



Levin, 2014

От сайтов связывания факторов транскрипции к системной биологии

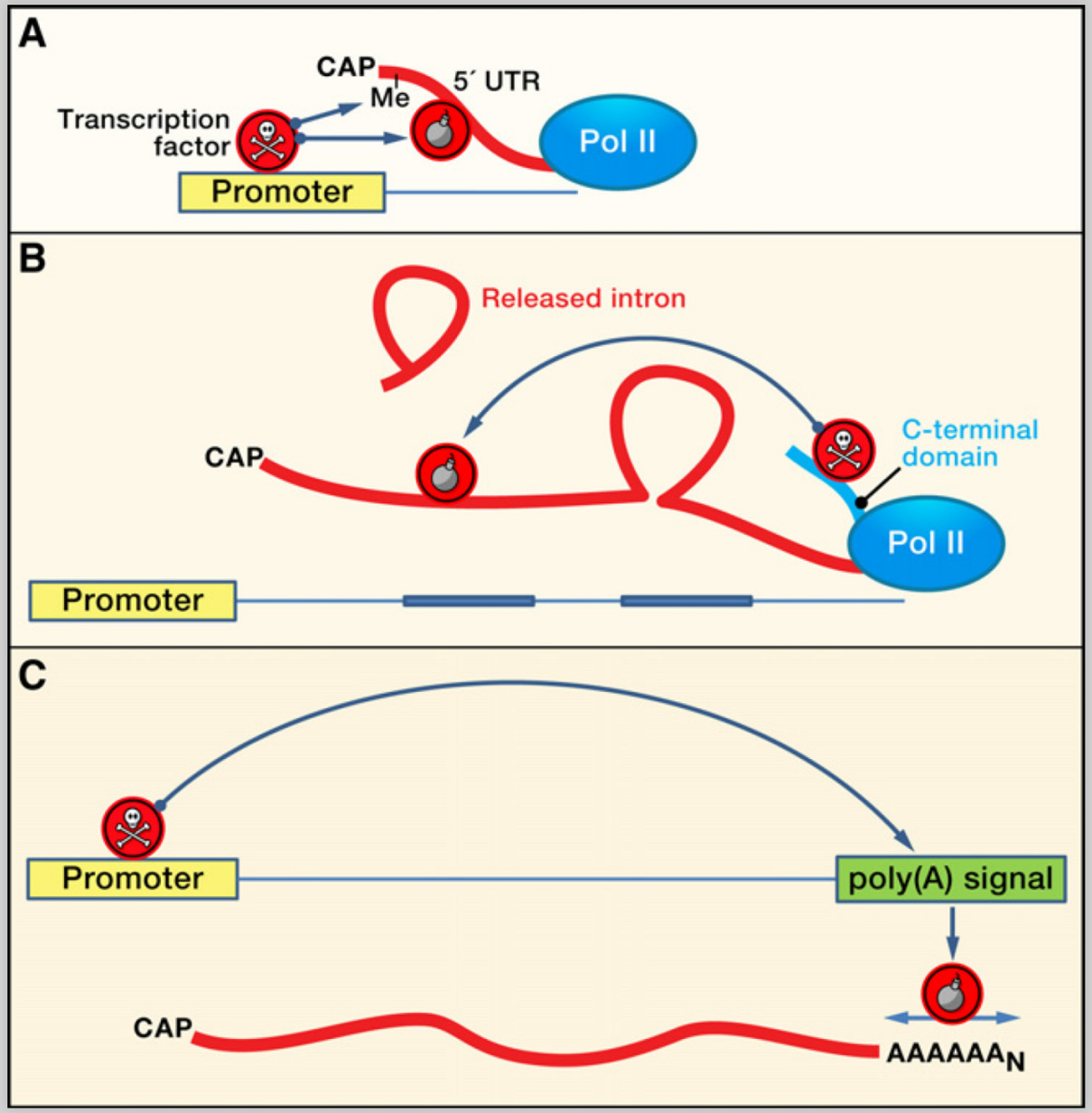




Transcription therefore is not only a catalyst of mRNA synthesis but also provides a platform that enables imprinting, which coordinates between transcription and mRNA decay.
Haimovich et al., Biochim Biophys Acta, 2013

Promoter sequences direct cytoplasmic localization and translation of mRNAs during starvation in yeast.
Zid and O'Shea, Nature, 2014

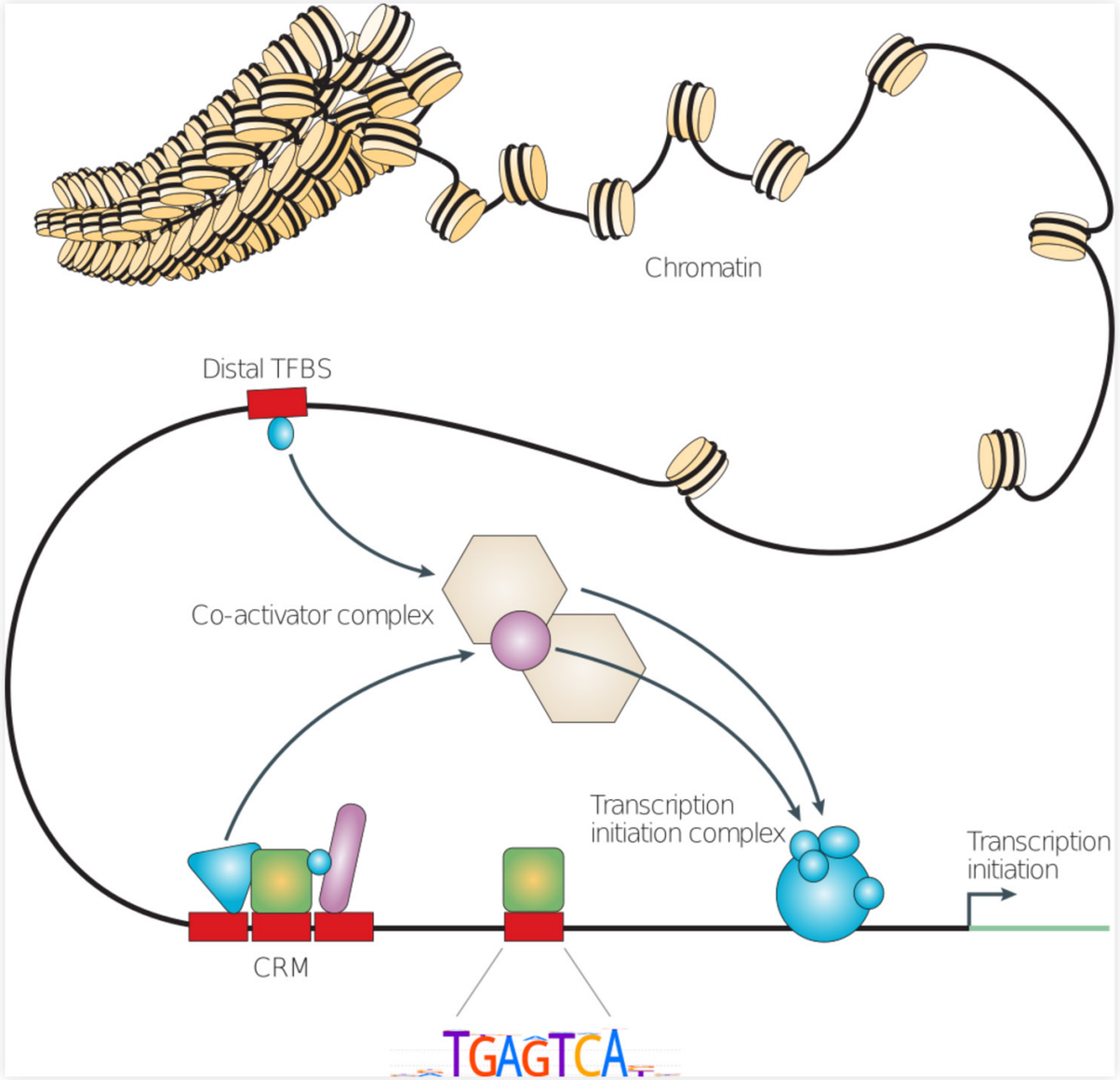
Как транскрипция может определять судьбу мРНК



Belofatto, 2011

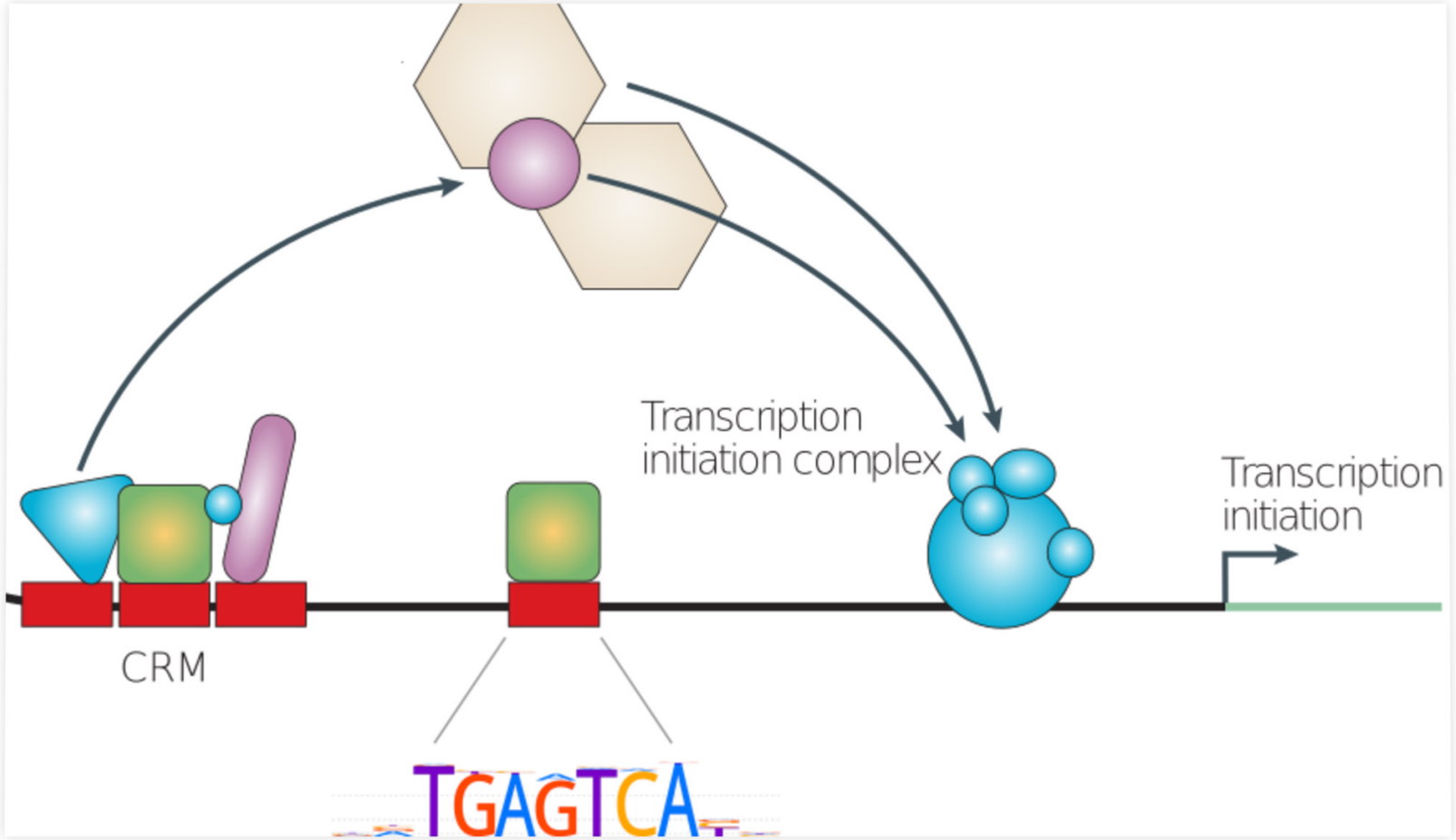
Вернемся к ДНК-сайтам связывания

Wasserman & Sandelin, 2004

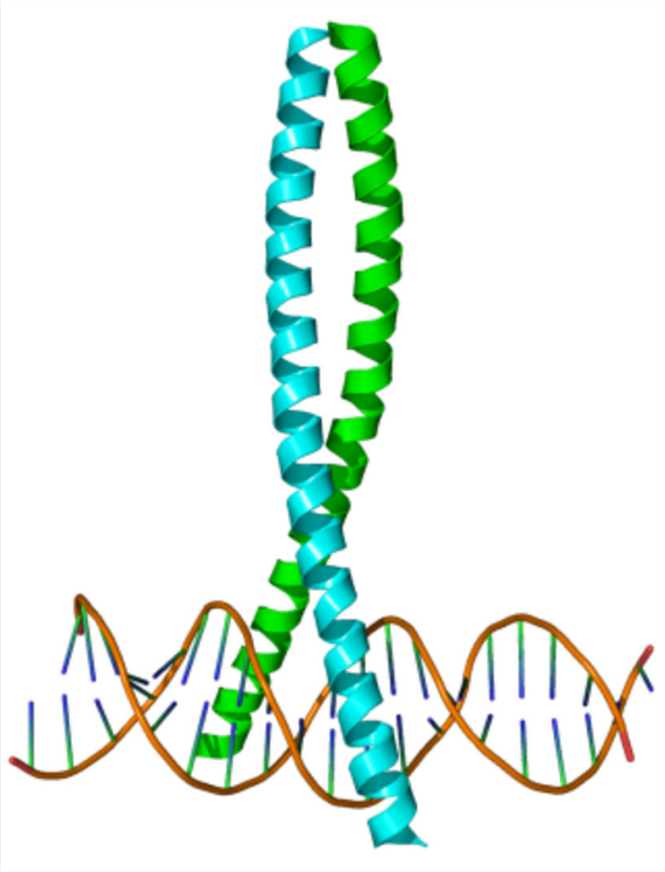


Вернемся к ДНК-сайтам связывания

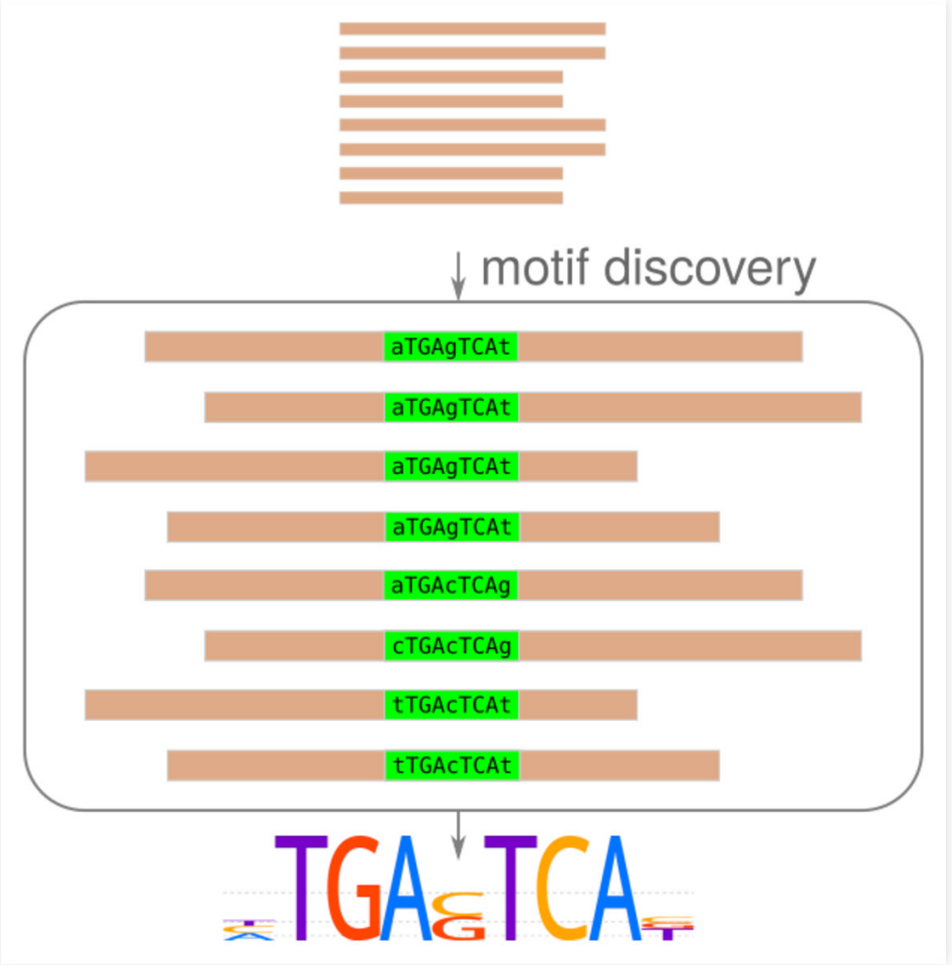
Wasserman & Sandelin, 2004



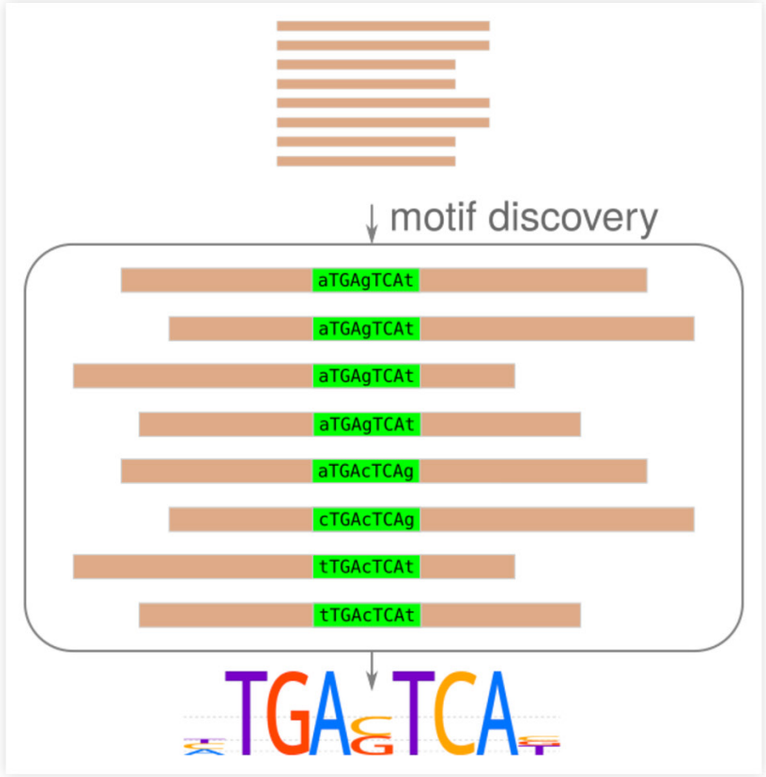
ДНК-белковое узнавание

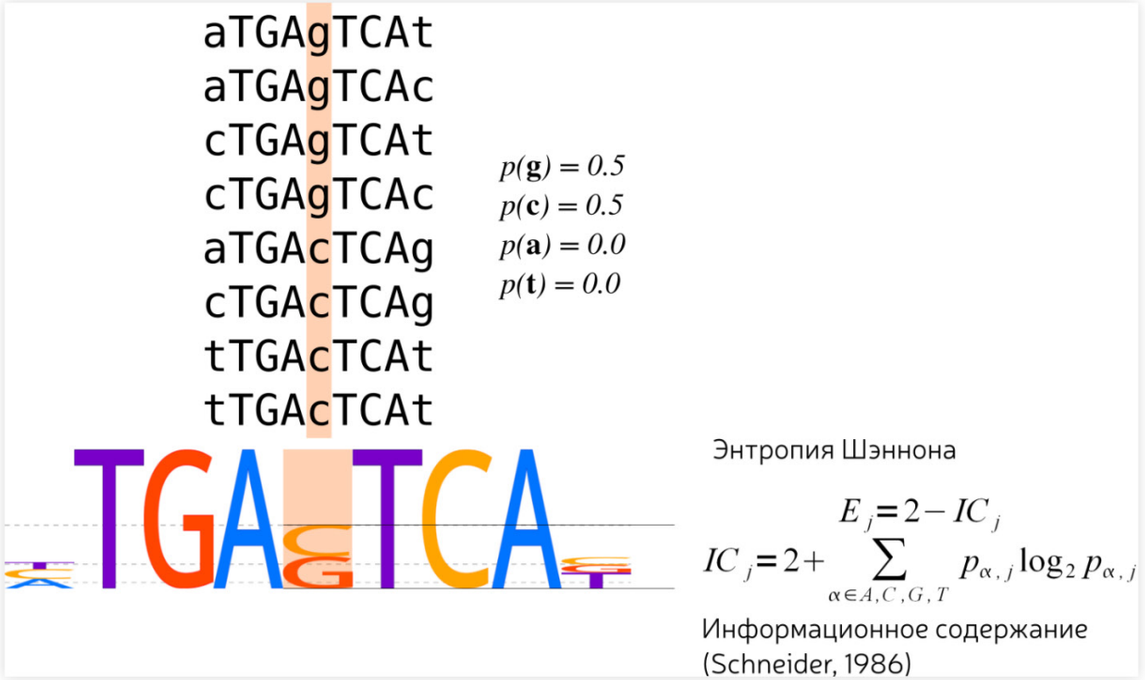


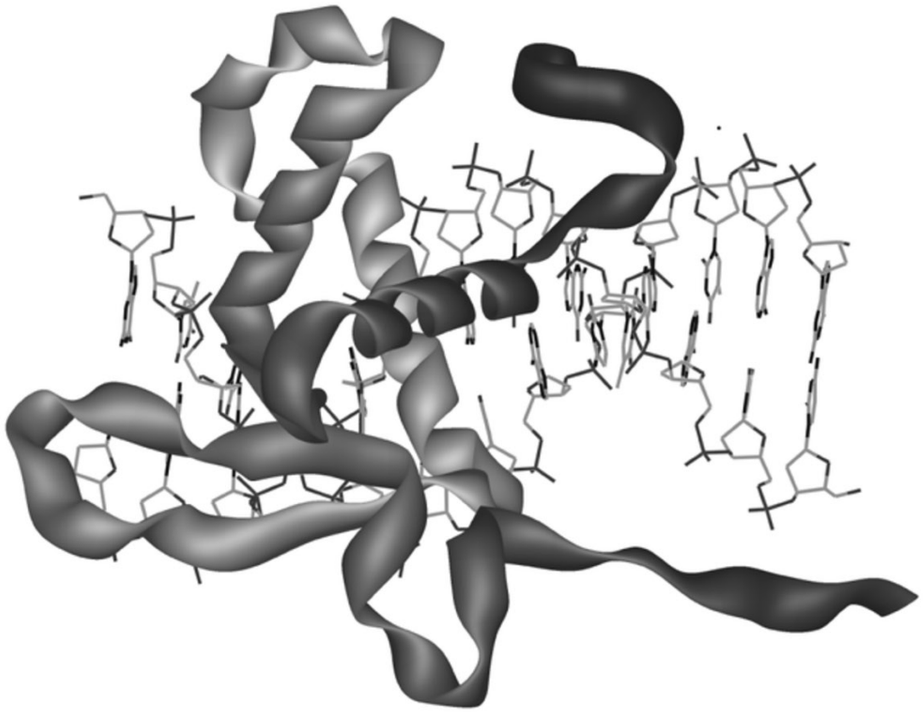
ДНК-белковое узнавание

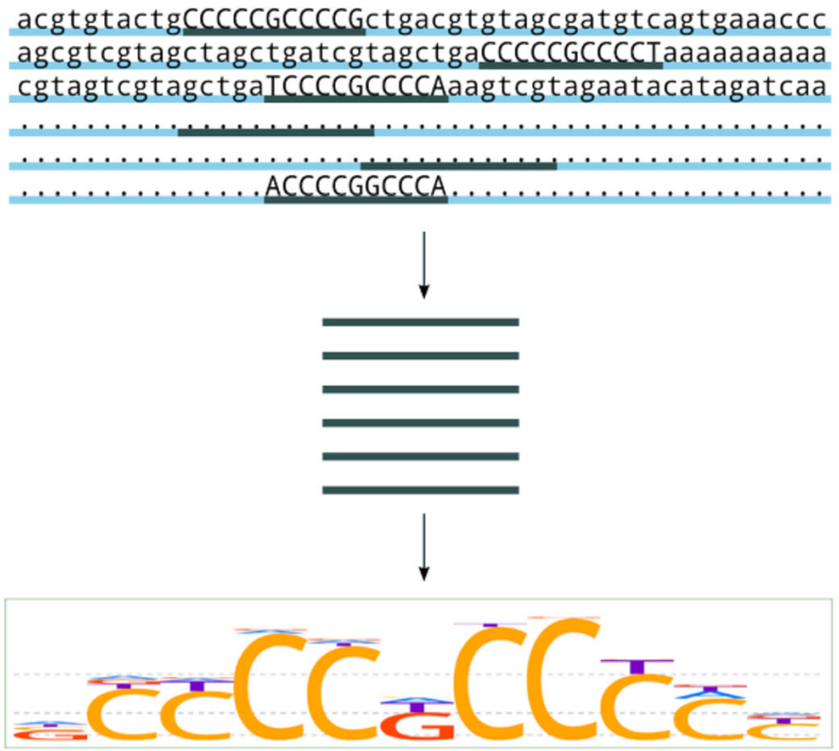


ДНК-белковое узнавание







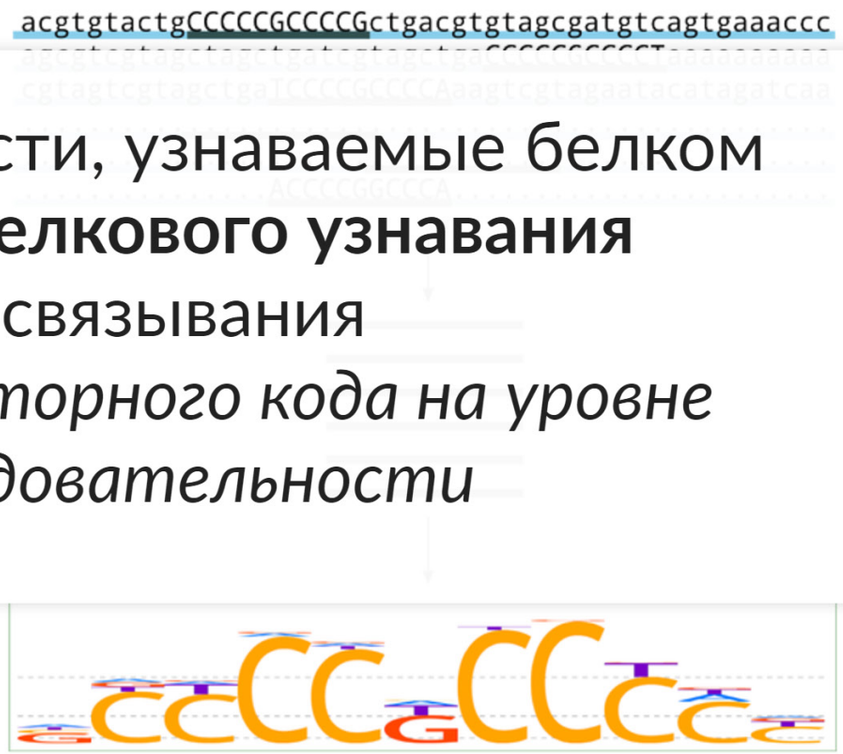

acgtgtactgCCCCCGCCCCGctgacgtgtagcgatgtcagtgaaaccc
agcgtcgtagctagctgatcgtagctgaCCCCCGCCCCTaaaaaaaaaa
cgtagtcgtagctgaTCCCCGCCCCAaagtcgtagaatacatagatcaa
ACCCCGGCCCA

Известные последовательности, узнаваемые белком

=> **паттерн-мотив ДНК-белкового узнавания**

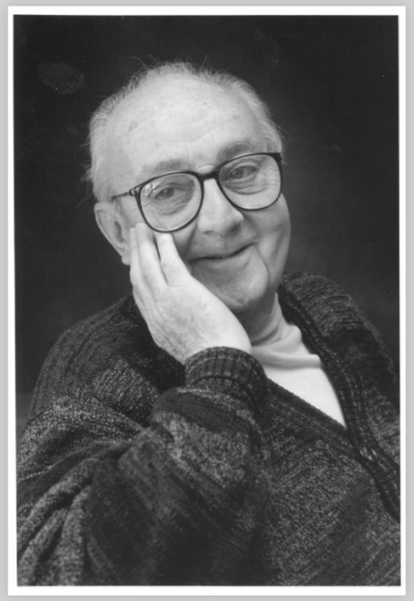
= модель сайтов связывания

=> базовый элемент *регуляторного кода на уровне нуклеотидной последовательности*



George Edward Pelham Box

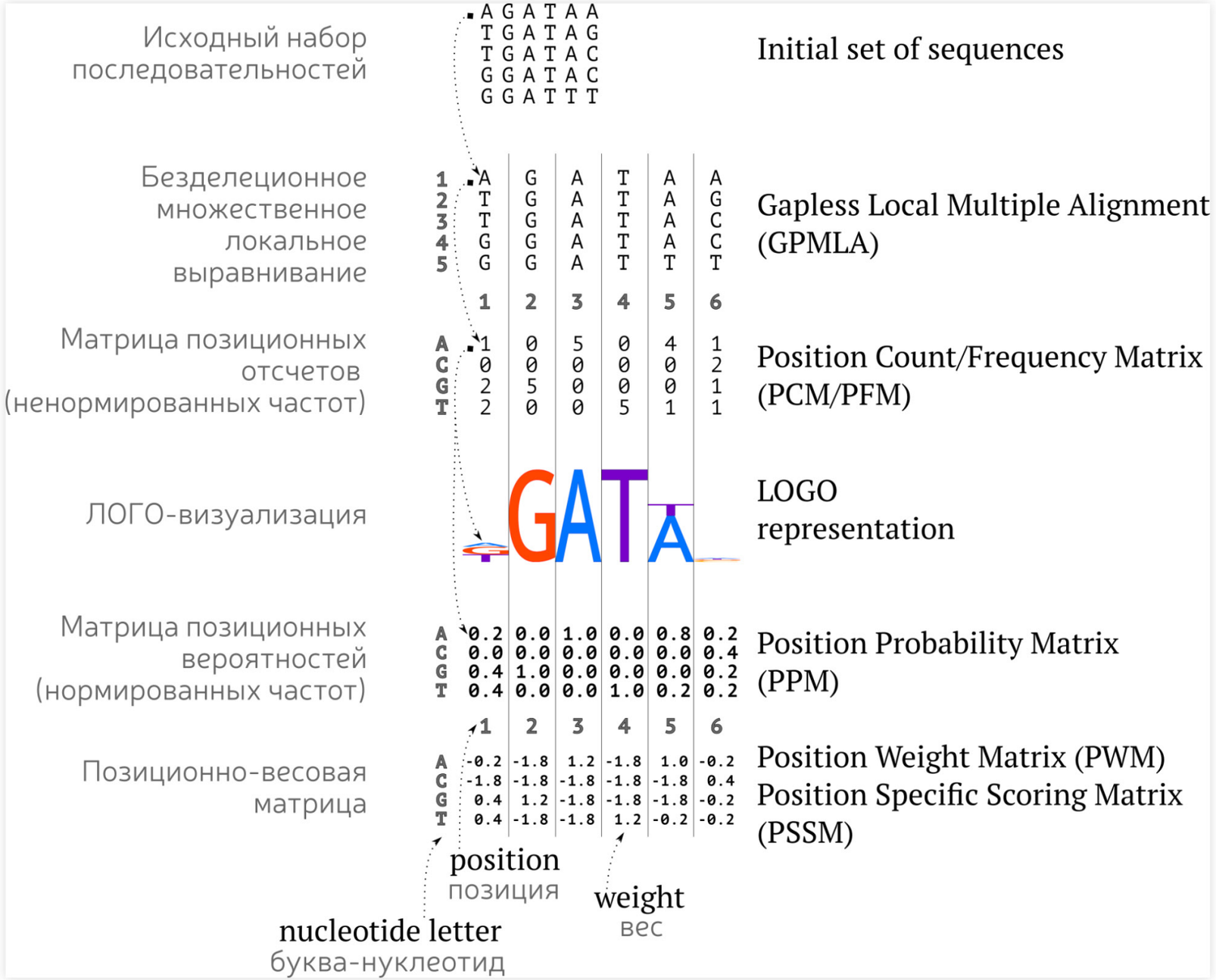
(October 18, 1919 – March 28, 2013)



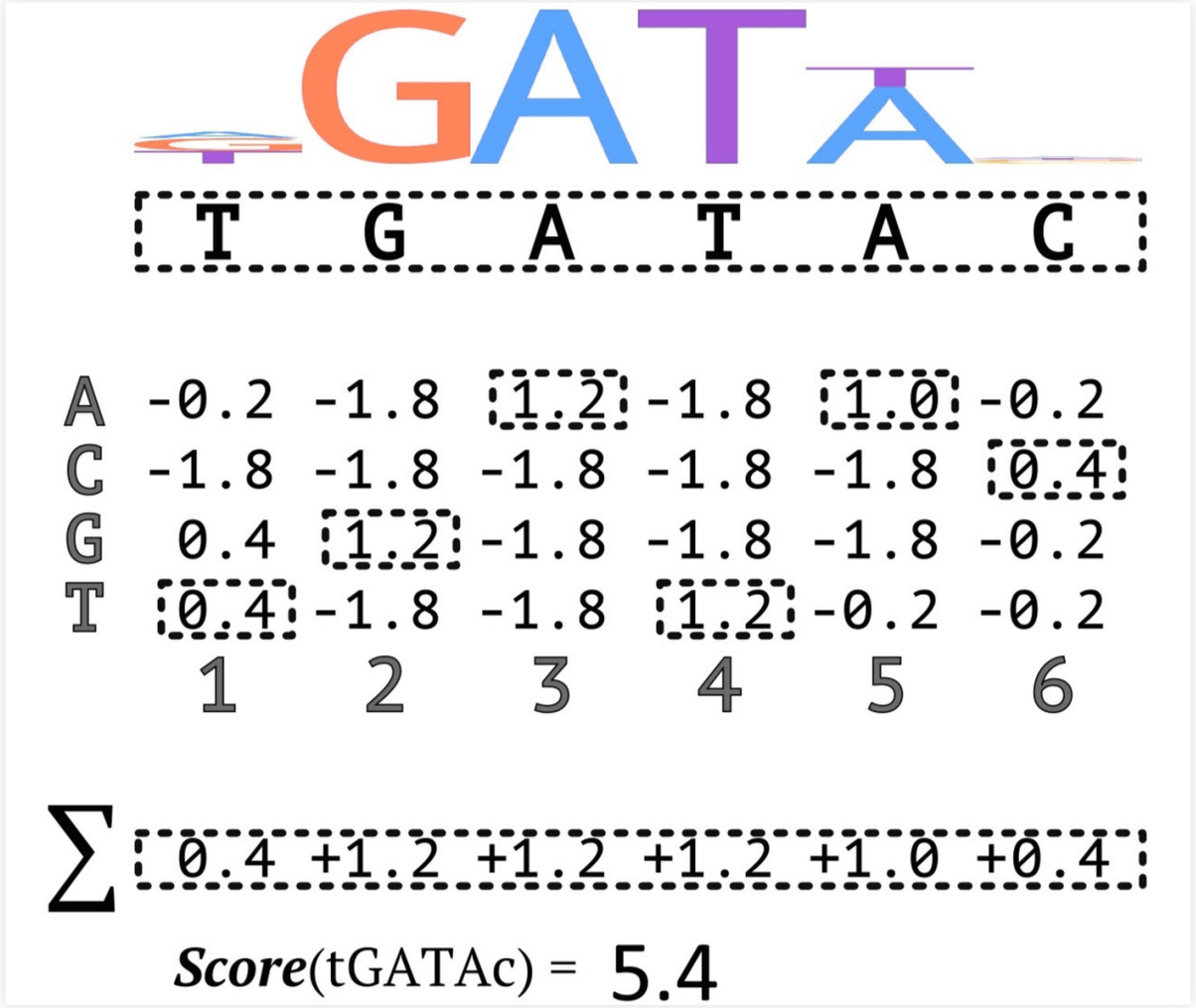
Essentially, all models are wrong, but some are useful.

in *Empirical Model-Building and Response Surfaces (1987)*

Позиционно-весовая матрица как модель мотива



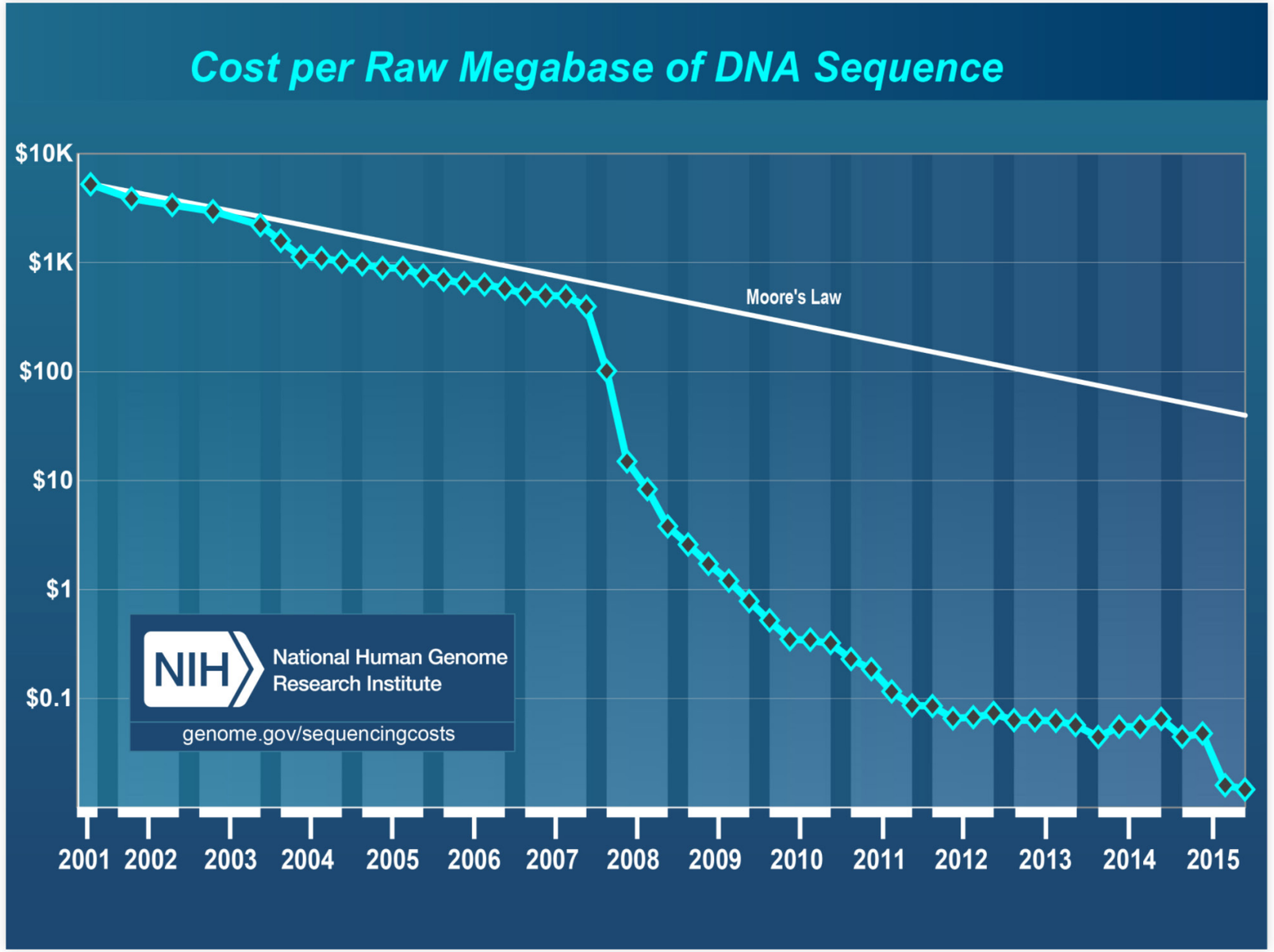
Позиционно-весовая матрица как модель мотива



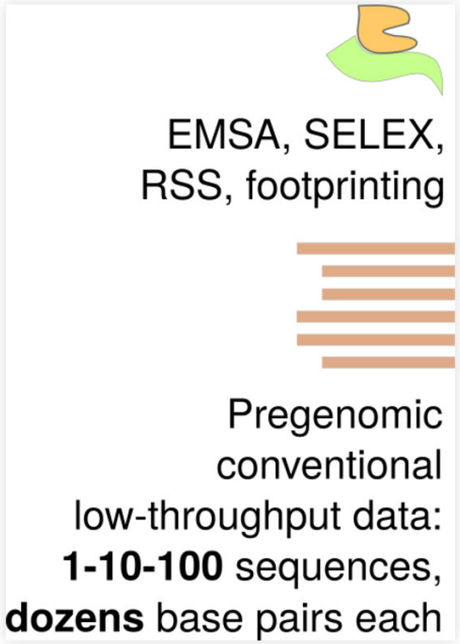
Berg, von Hippel, 1986-1988

Легко ли искать мотивы в реальных экспериментальных данных по ДНК-белковому узнаванию?

Microprocessor Transistor Counts 1971-2011

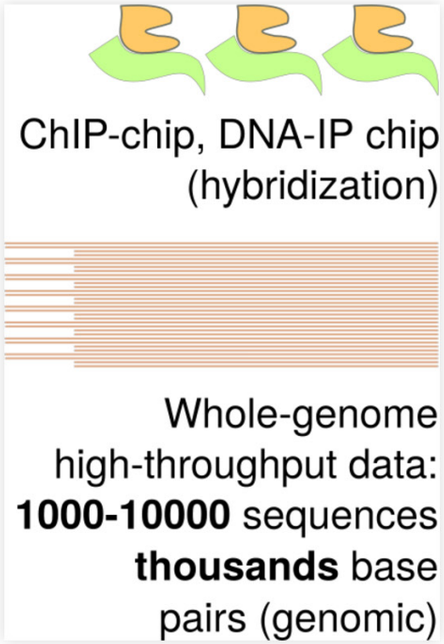

Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

Эволюция экспериментальных методов по изучению ДНК-белкового узнавания



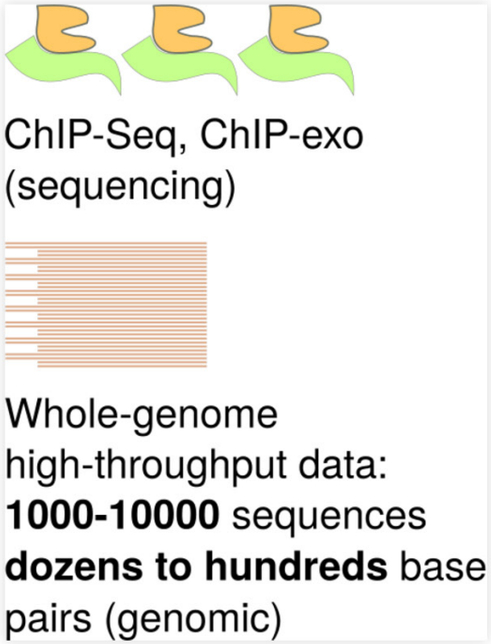
EMSA, SELEX, RSS, footprinting

Pregenomic conventional low-throughput data: **1-10-100** sequences, **dozens** base pairs each



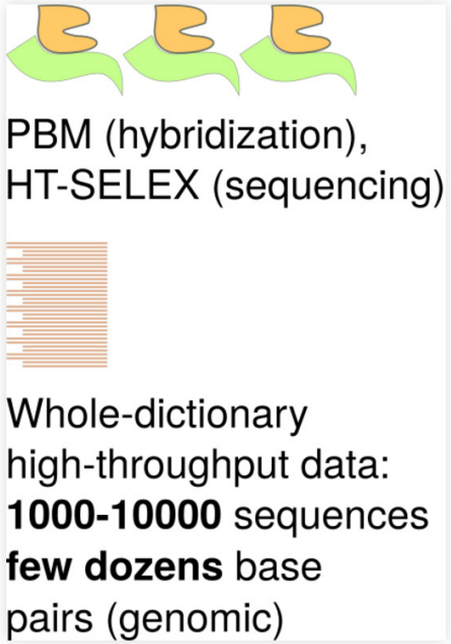
ChIP-chip, DNA-IP chip (hybridization)

Whole-genome high-throughput data: **1000-10000** sequences **thousands** base pairs (genomic)



ChIP-Seq, ChIP-exo (sequencing)

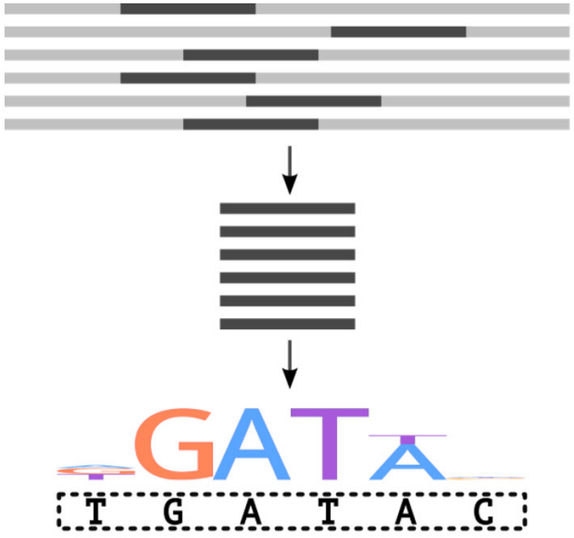
Whole-genome high-throughput data: **1000-10000** sequences **dozens to hundreds** base pairs (genomic)



PBM (hybridization), HT-SELEX (sequencing)

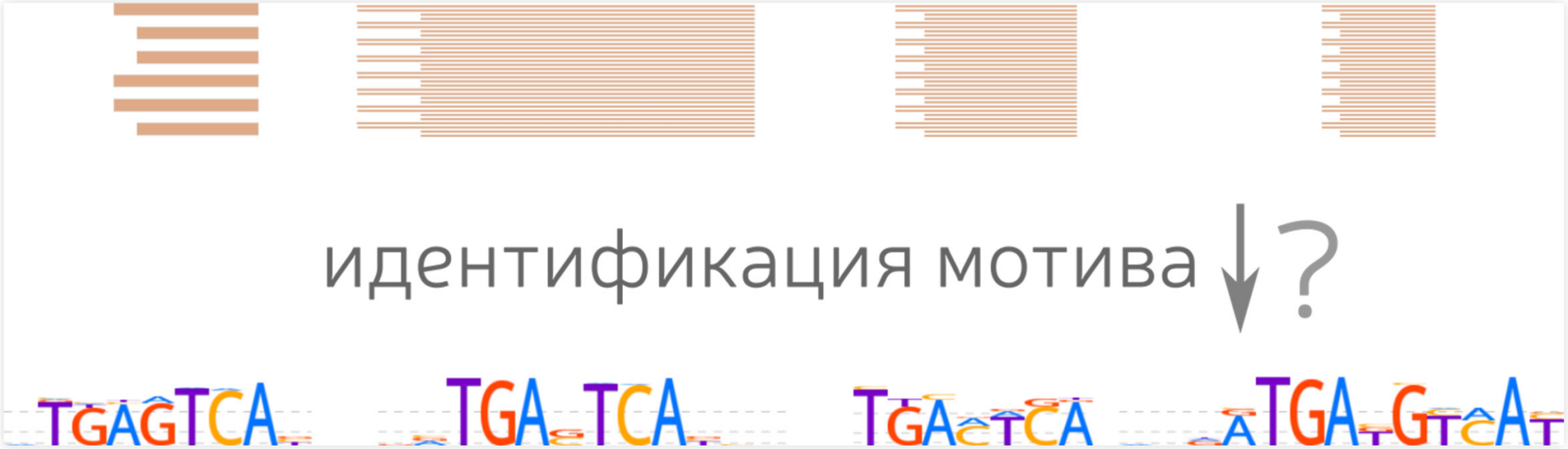
Whole-dictionary high-throughput data: **1000-10000** sequences **few dozens** base pairs (genomic)

Идентификация мотивов: любимая задача биоинформатиков

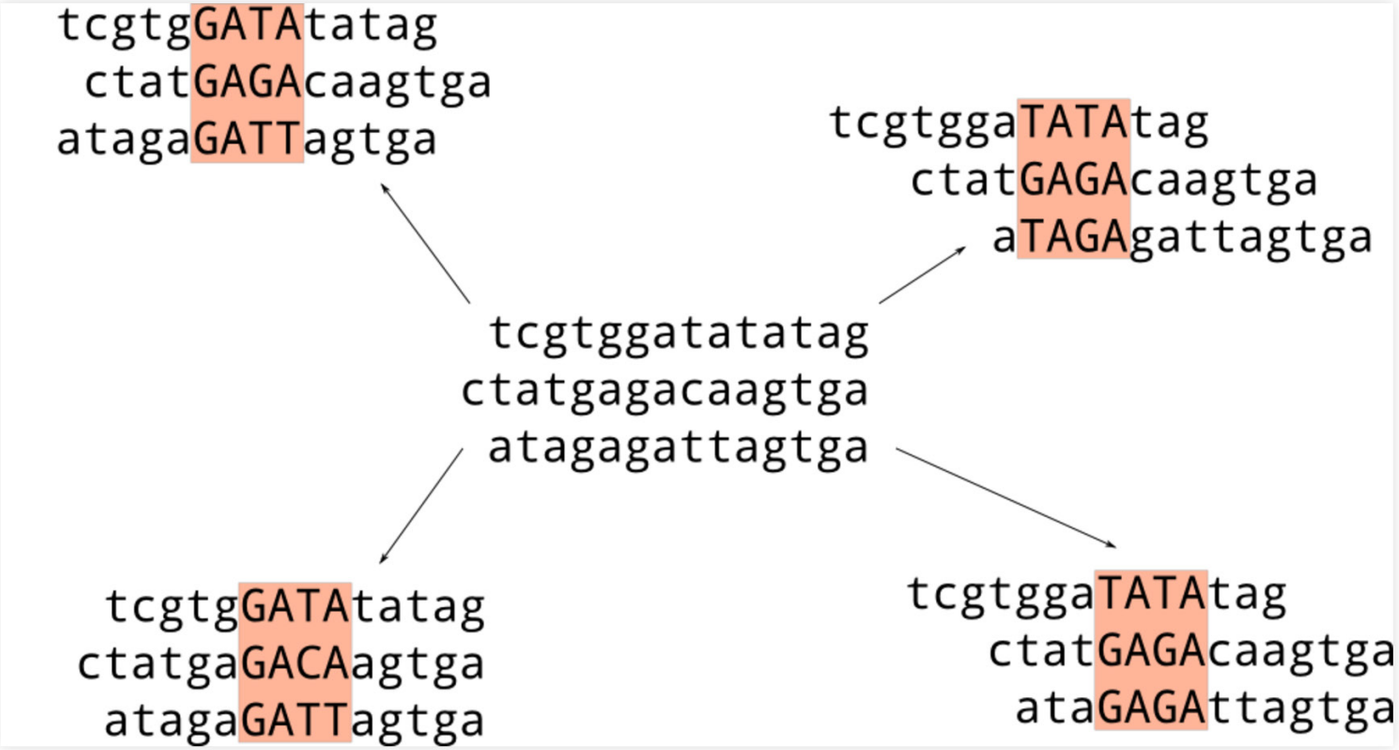


WordUP
Gibbs sampler
MACAW MEME AlignACE
Oligo-Analysis Consensus Dyad-Analysis
WINNOWER ANN-Spec SMILE Verbumculus
MobyDick YMF Bioprospector Co-Bind ITB Weeder
MotifSampler MITRA MDScan
Projection Footprinter MOPAC DMotif
PhyloCon LOGOS EC GLAM Improbizer QuickScore
SeSiMCMC PhyME OrthoMEME FMGA PHYLONET PhyloGibbs
GIMF WordSpy MaMF EMD GibbsST
MUSA GAME ALSE MotifSeeker GRISOTTO PhyloScan
HMS Bigfoot HOMER DMF
SITECON SiteGA and hundreds more...

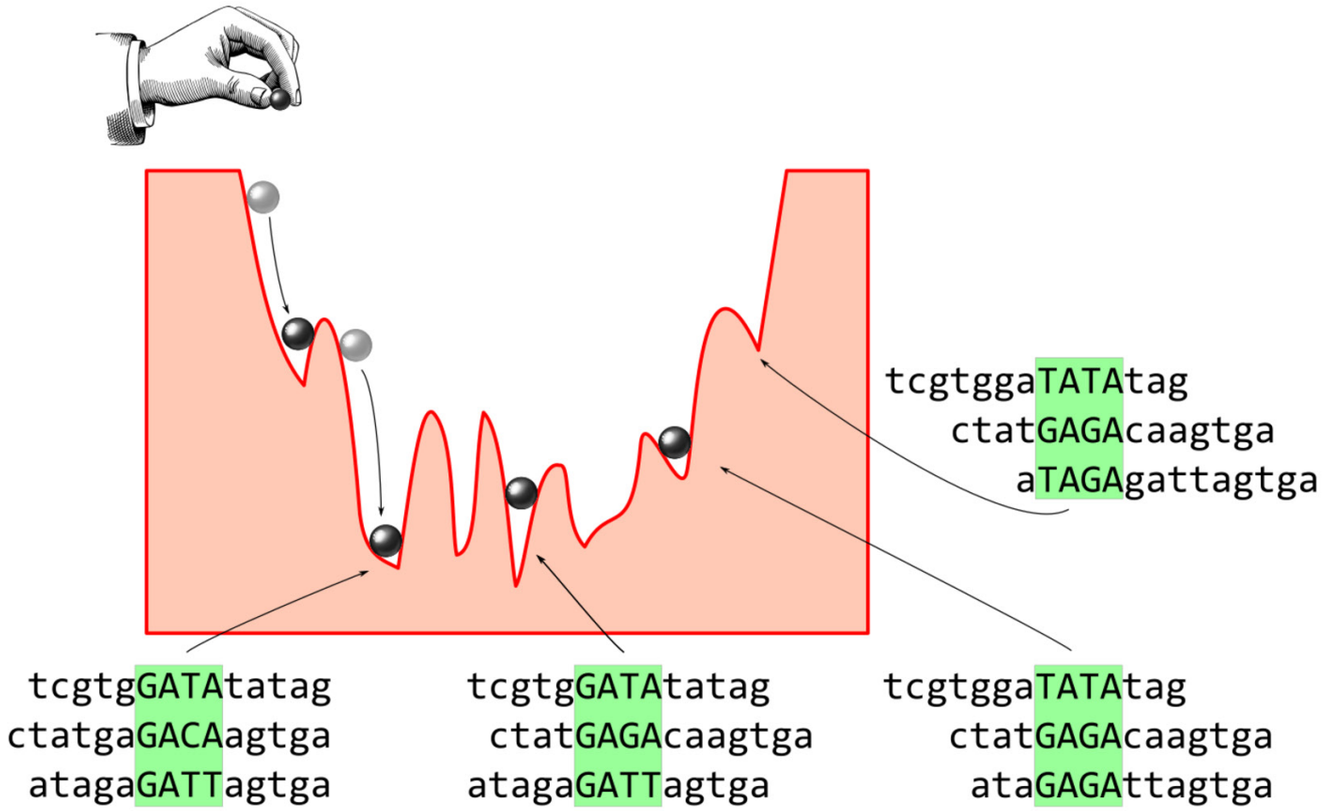
Содержательная проблема: несогласованность результатов



Корректность задачи идентификации мотива

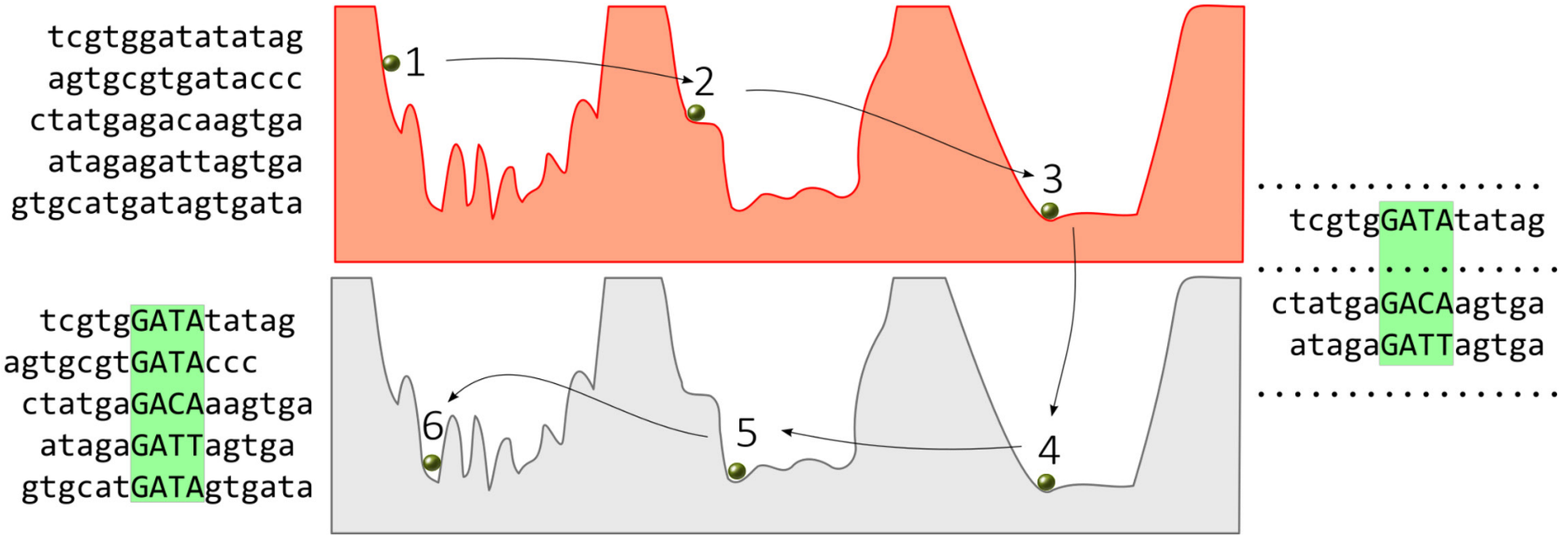


Идентификация мотивов



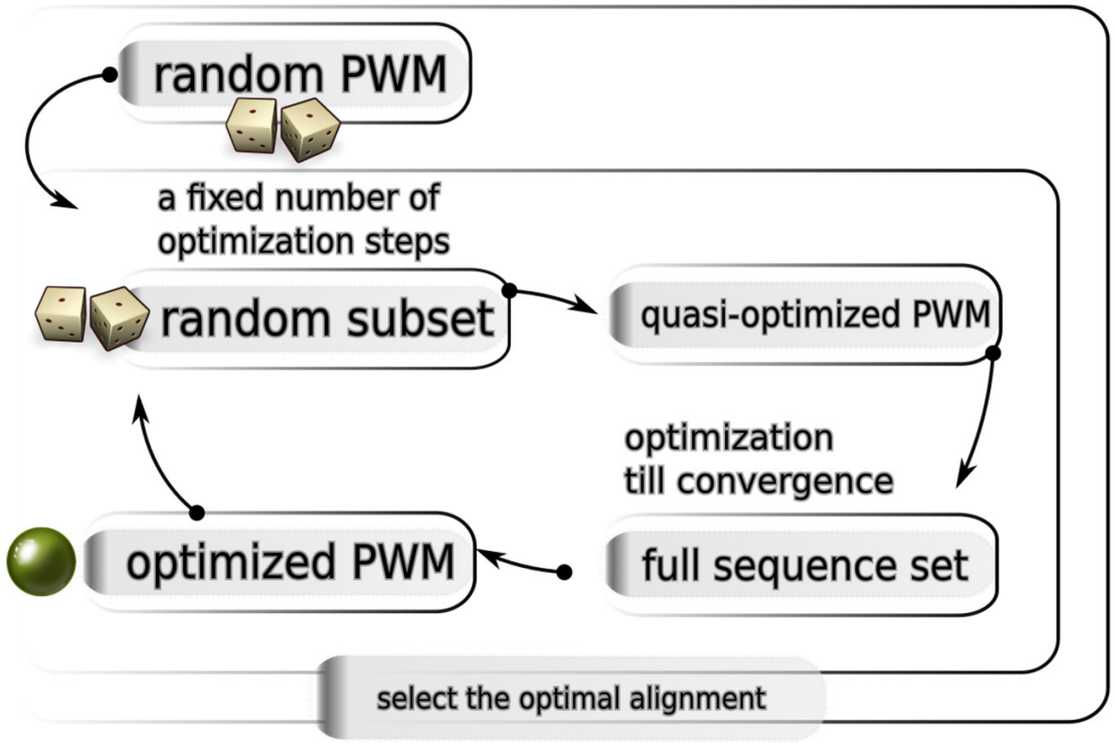
Kulakovskiy et al., 2010, 2013

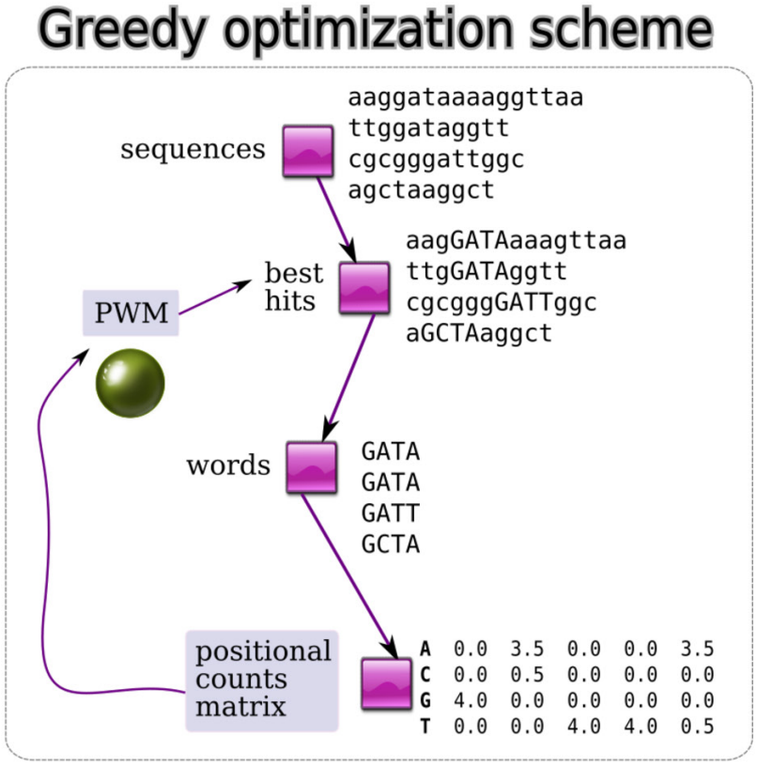
Идентификация мотивов



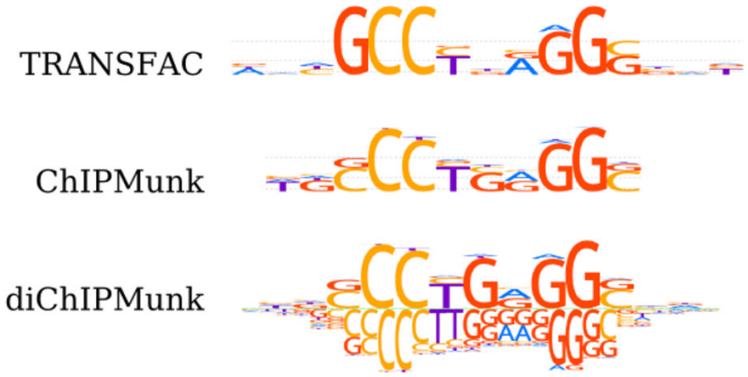
Kulakovskiy et al., 2010, 2013

Идентификация мотивов





Учет зависимостей между нуклеотидами (динуклеотиды)

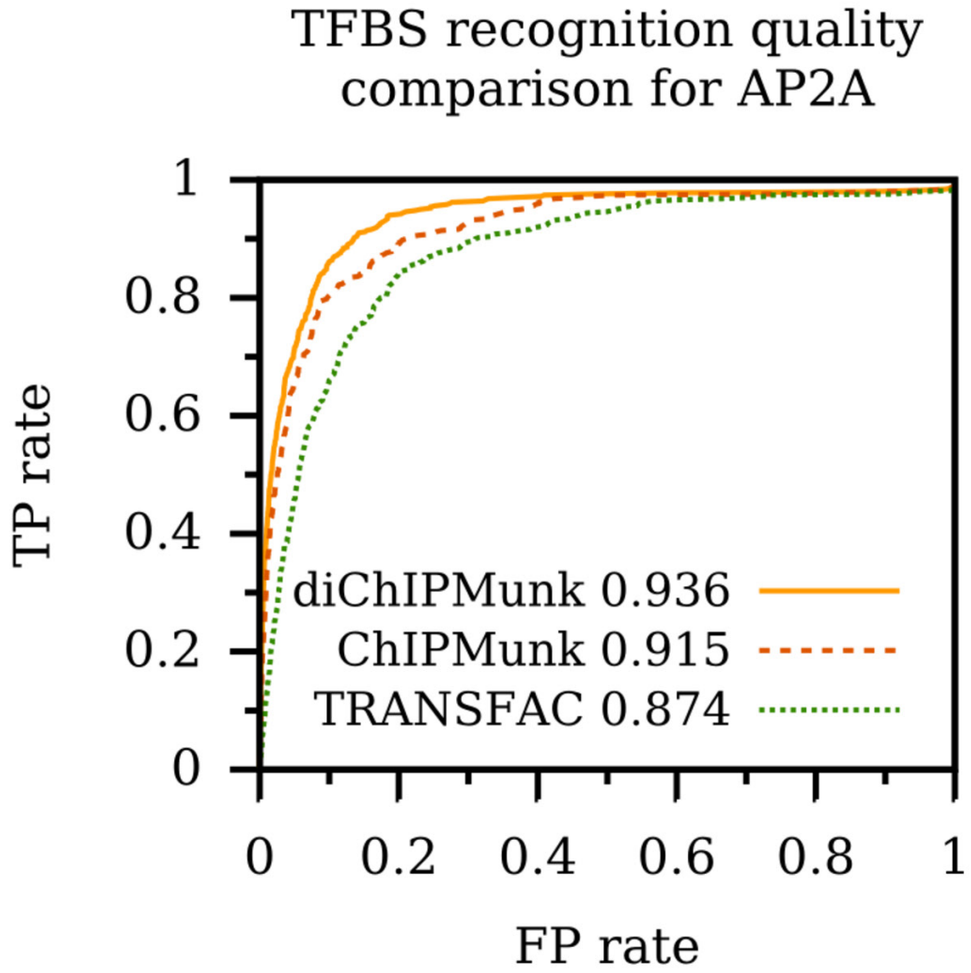


GATA:

G'A'T'A or GA'AT'TA

mononucleotide alphabet {A,C,G,T}

dinucleotide superalphabet {AA,AC,...TT}



Чего мы не сделали?

GPGPU-реализация

или эффективный поиск вхождений на этапе оптимизации

"тринуклеотидные" мотивы?

Что уже сделано:

Generalized dinucleotide PWMs

Remote dependency models

Variable-order models

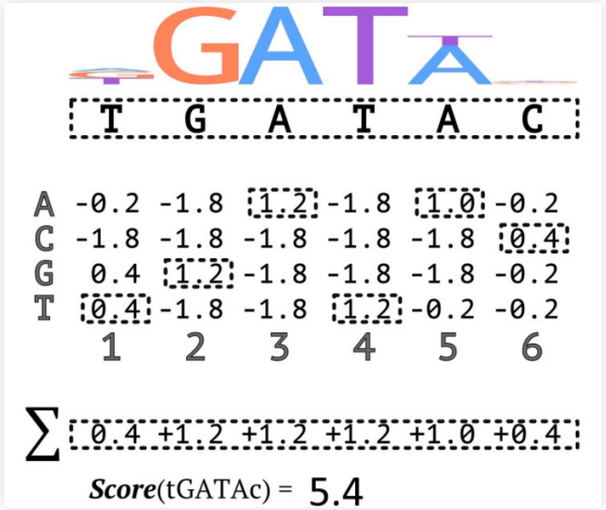
DeepBind (k-mers)

Настоящие и ненастоящие мотивы связывания

Оценки "скоры" весовых матриц

Что хорошо: легко считать, чем-то похожи на энергию

Что плохо: *в попугаях я длиннее*, предположение об аддитивности



От скоров-оценок к P-значениям

Motif model (e.g. positional weight matrix, PWM)

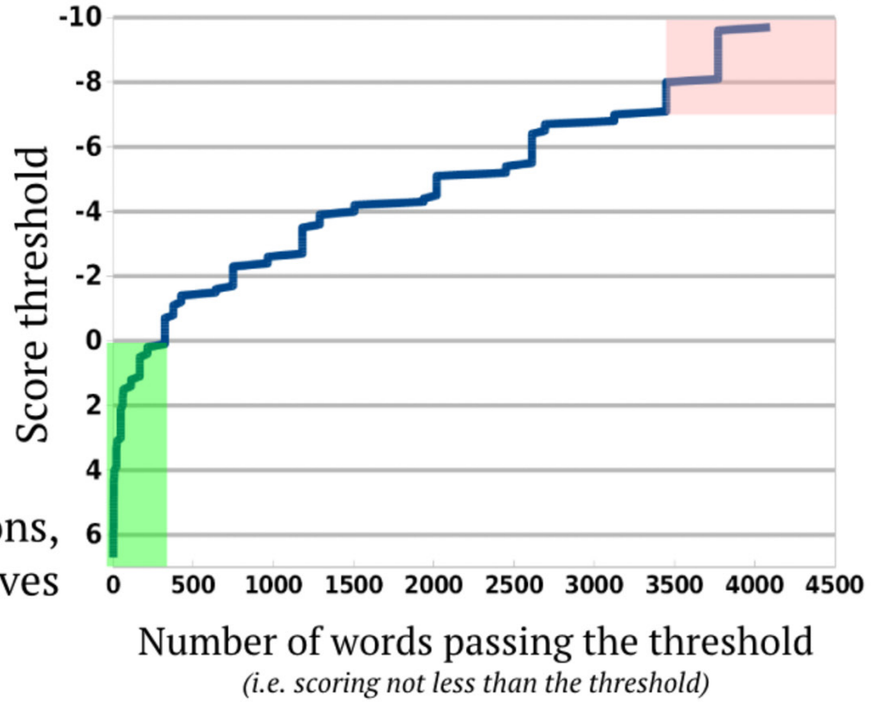
	1	2	3	4	5	6
A	-1.6	-1.6	0.96	-1.6	-1.6	0.96
C	-1.6	-1.6	0.00	-1.6	-1.6	-1.6
G	1.22	1.22	-1.6	-1.6	-1.6	-1.6
T	-1.6	-1.6	-1.6	1.22	1.22	0.00

PWM
GGATTA → S_{GGATTA}=1.22+1.22+0.96+1.22+1.22+0.96=**6.8** the best score

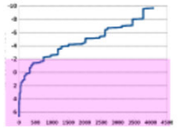
S_{GGGGGG}=2.44-6.4=**-3.96**

S=**-9.6** the worst score

more predicted TFBS,
more false positive predictions



less TFBS predictions,
less true positives



Score threshold turns a motif model into a binary "yes/no" classifier!

От скоров-оценок к P-значениям

PWM: word > score

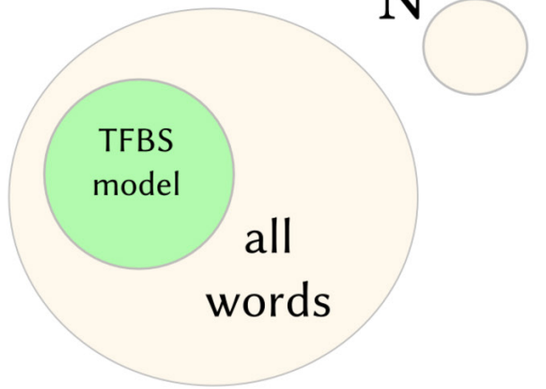
T G A T A C

	1	2	3	4	5	6
A	-0.2	-1.8	1.2	-1.8	1.0	-0.2
C	-1.8	-1.8	-1.8	-1.8	-1.8	0.4
G	0.4	1.2	-1.8	-1.8	-1.8	-0.2
T	0.4	-1.8	-1.8	1.2	-0.2	-0.2

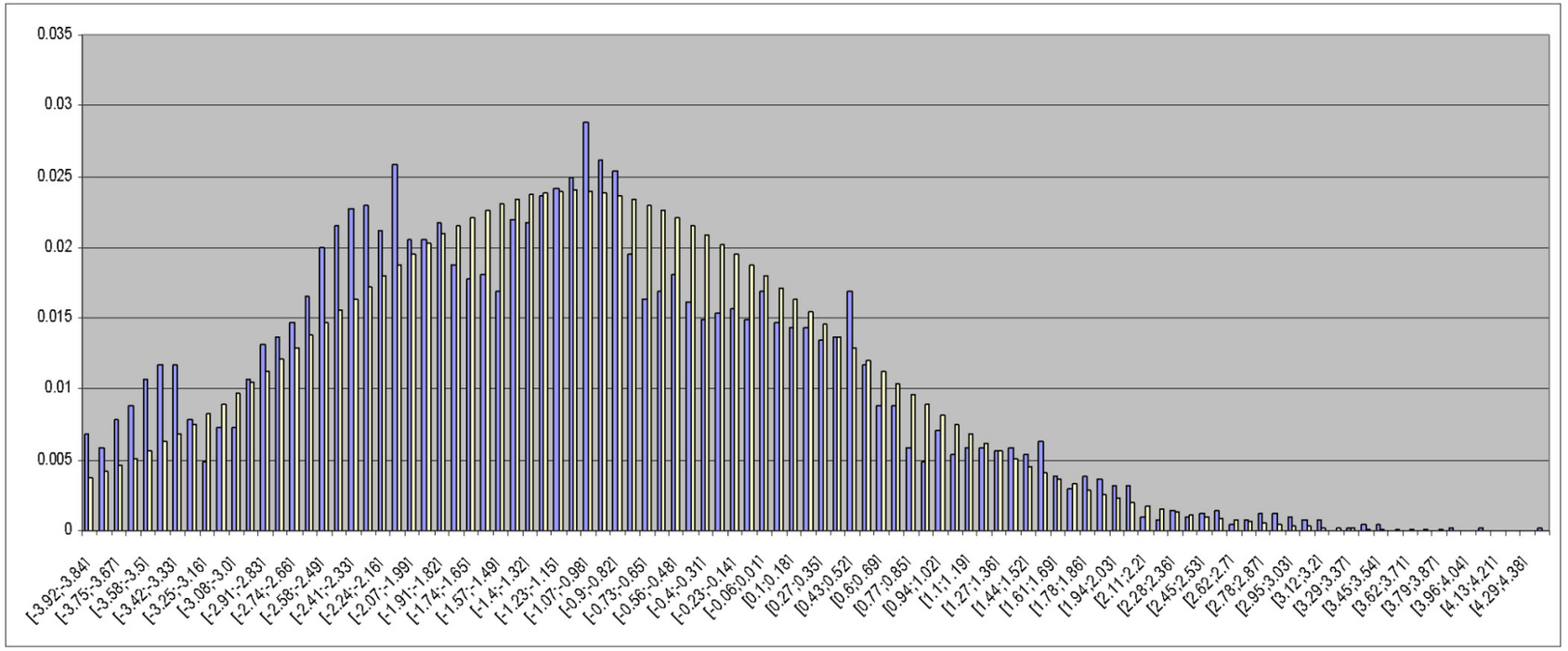
$\sum$ 0.4 +1.2 +1.2 +1.2 +1.0 +0.4

P-value: score > significance

$$P\text{-value} = \frac{N}{N}$$

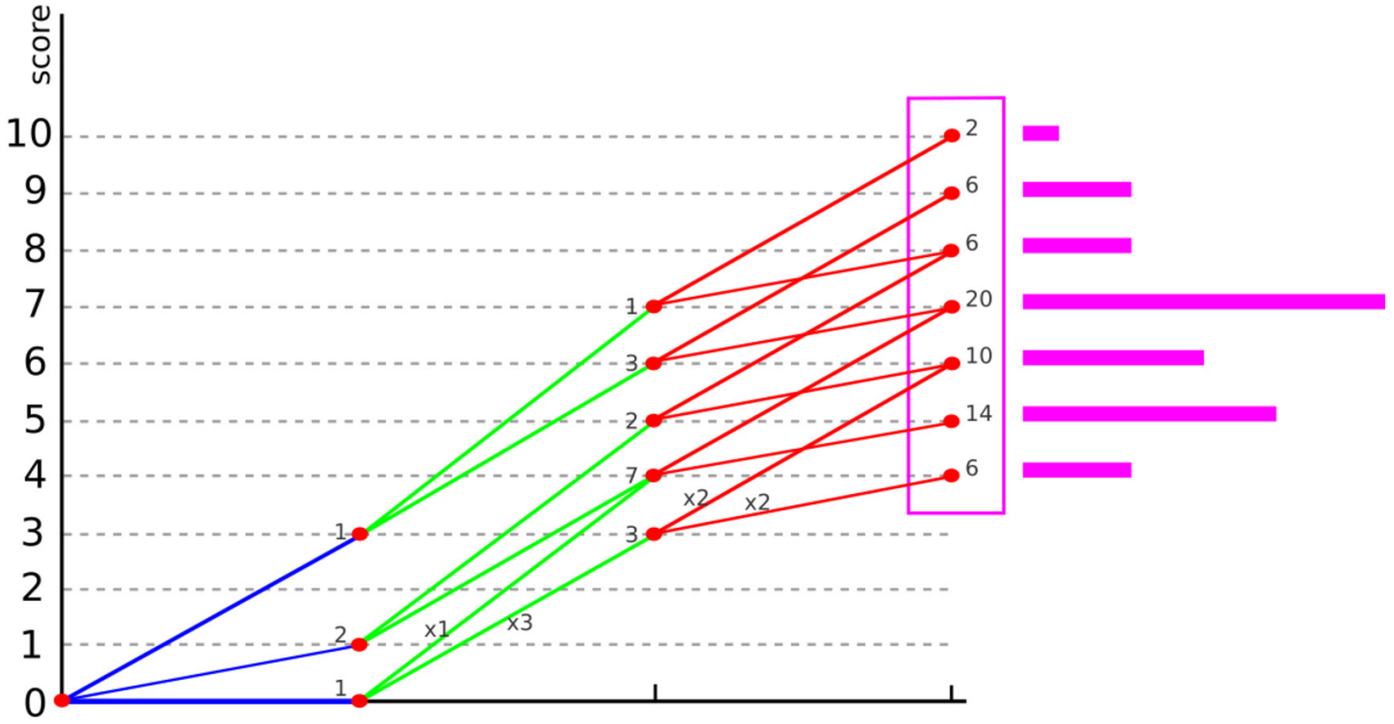


Как выглядит распределение возможных оценок



Уточненная оценка P-значений (площади хвоста)

A	1	3	1
C	1	3	3
G	0	4	3
T	3	3	1



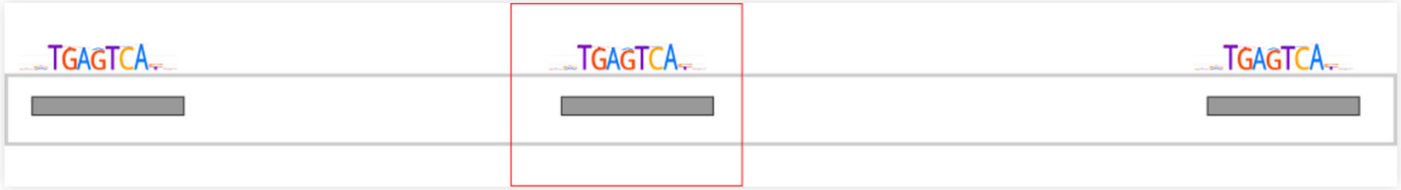
Достоверное предсказание для p=0.001

и отдельного олигонуклеотида

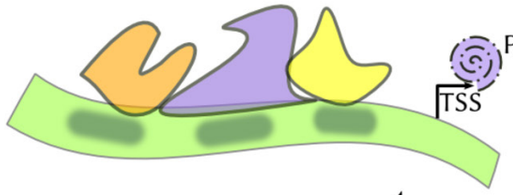
Существенно менее надежно - P~1

для последовательности длины 1000

Бессмысленно в полногеномном масштабе

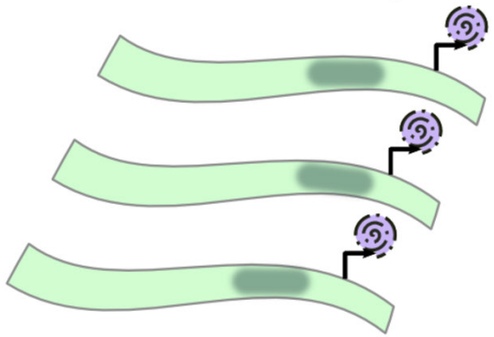


Практический анализ мотивов

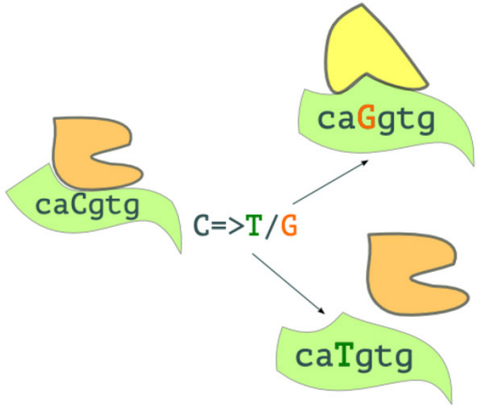


Позиционные предпочтения связывания,
композитные элементы

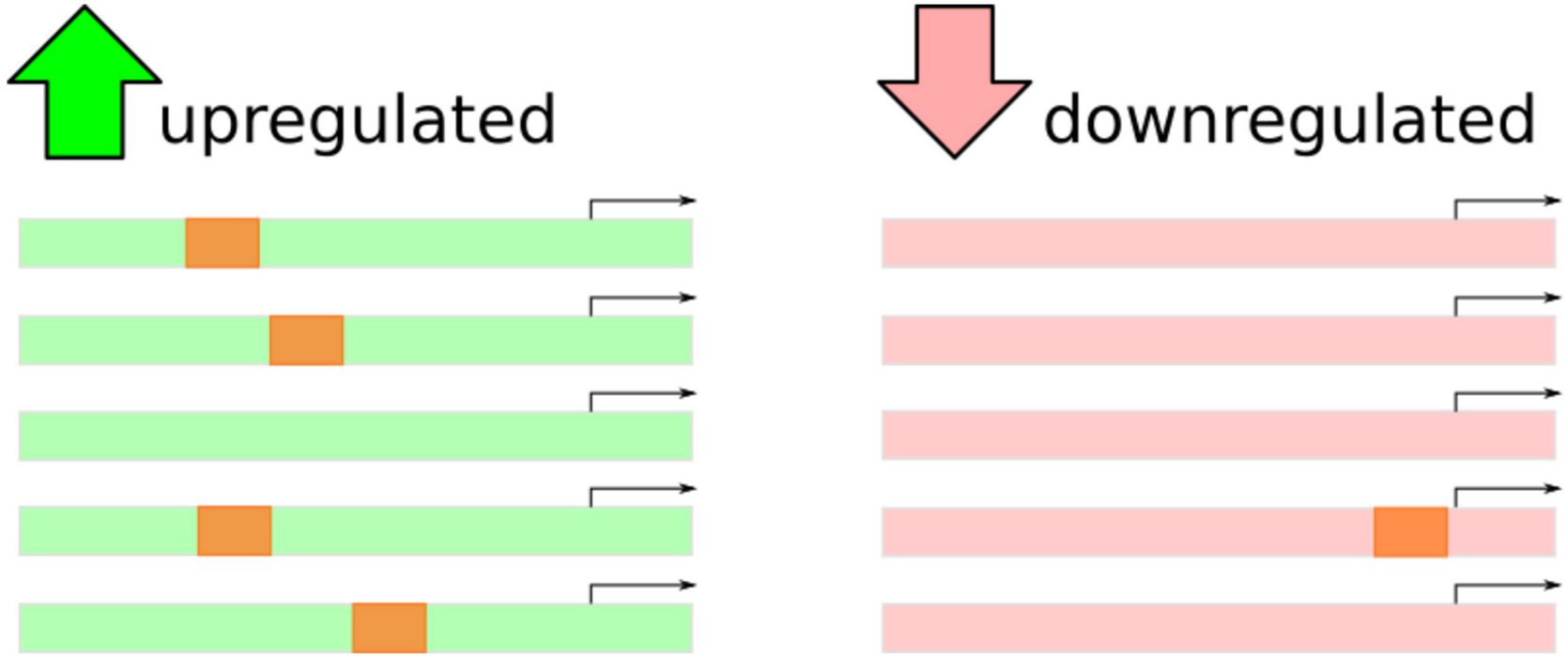
Сайты связывания
в регуляторных районах
зависимых генов



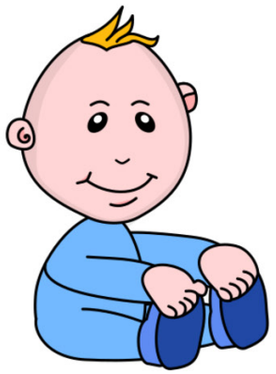
Функциональный эффект вариаций
в регуляторных районах



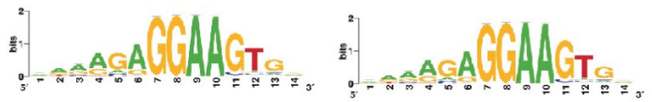
Предсказание сайтов связывания в промоторах дифференциально-экспрессируемых генов

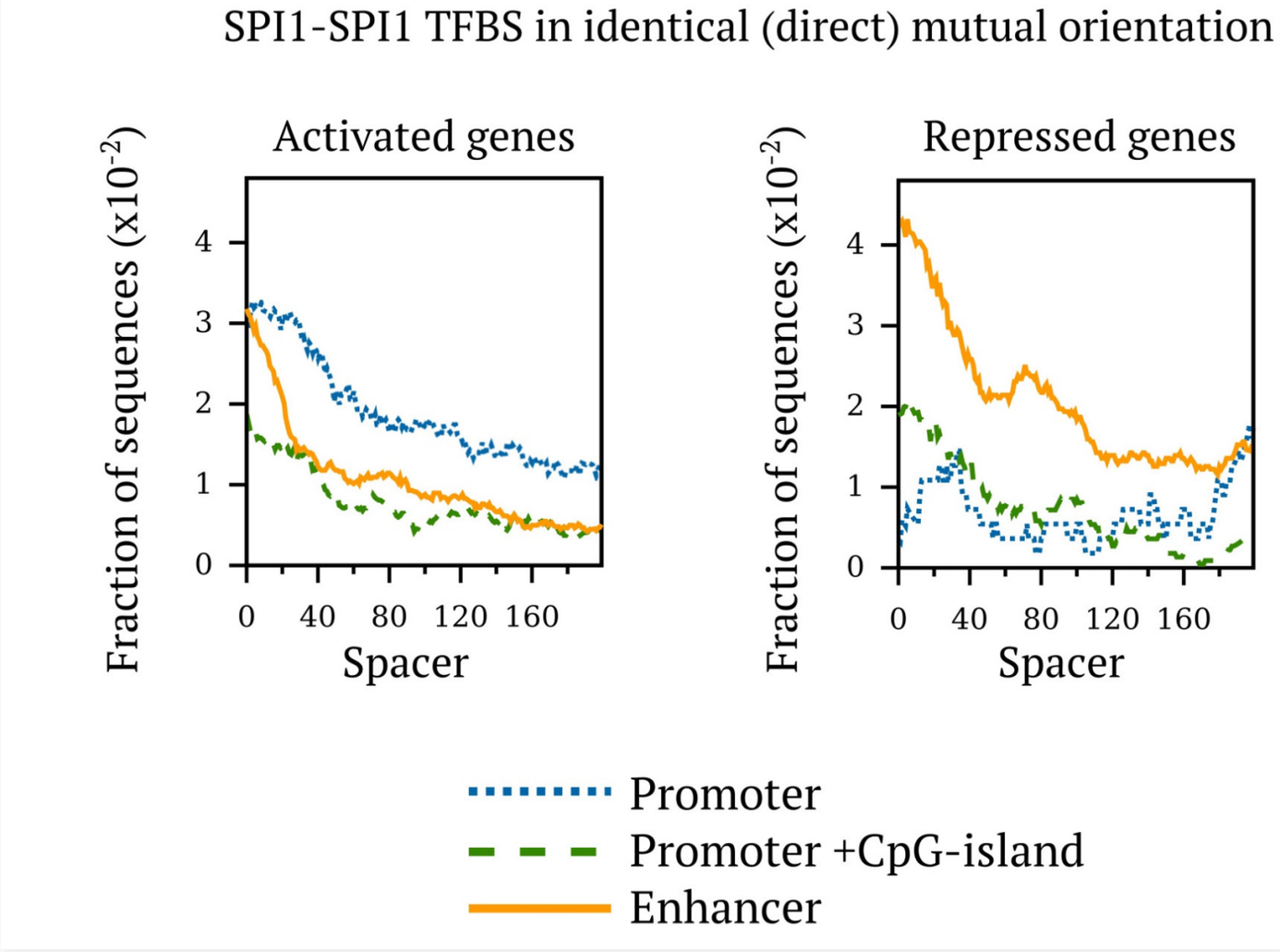


Поиск мотивов и любой тест на таблицах сопряженности

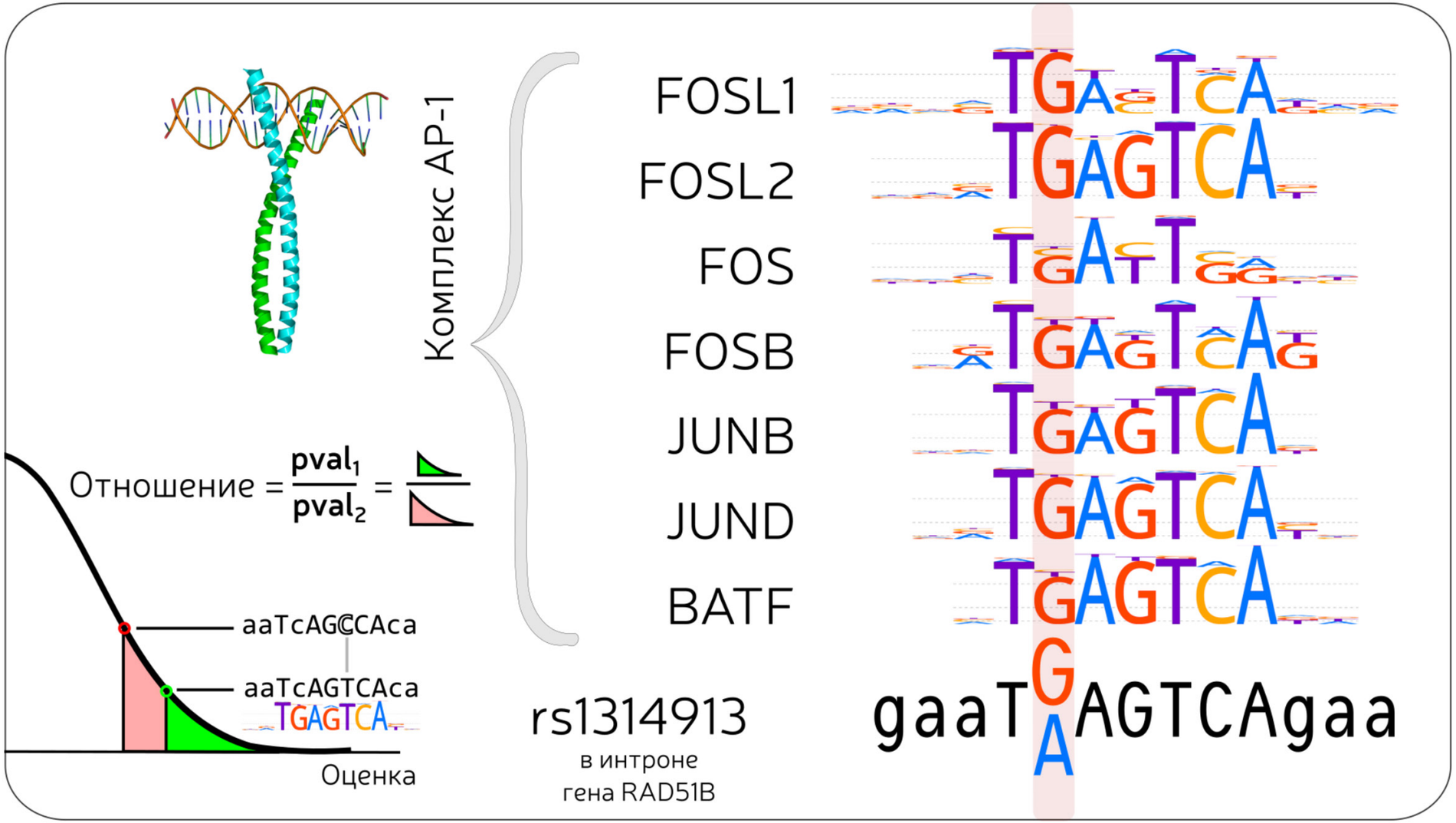
		
	4	1
	1	4

Альтернативный вариант: композитные элементы





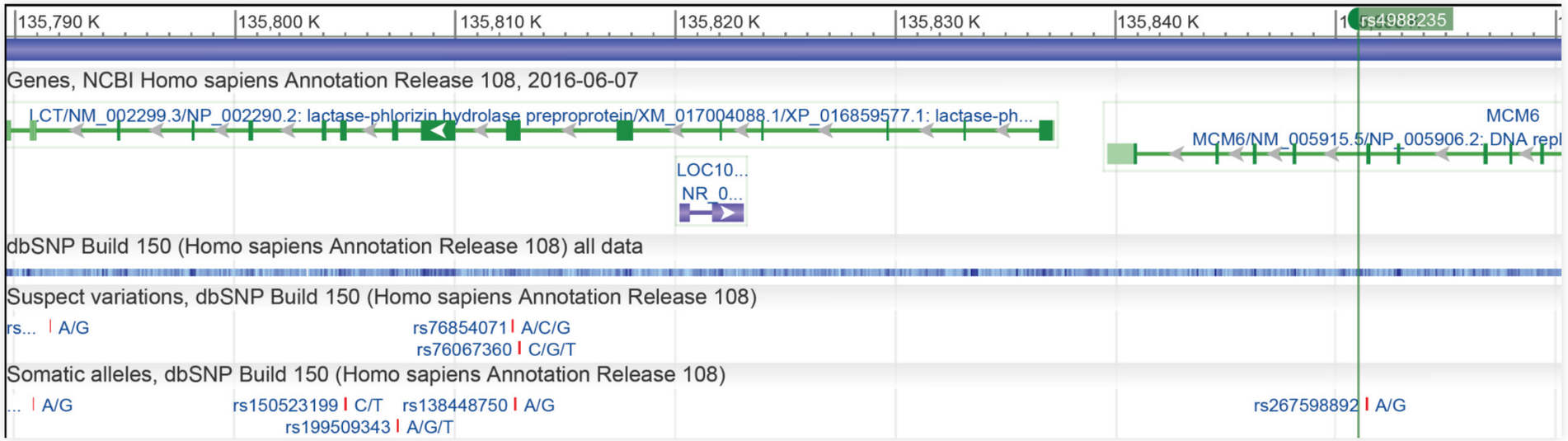
Оценка эффекта однонуклеотидных замен



Vorontsov et al., 2015

Доктор Леонард Хофстедтер

- Физик, кандидат наук
- Носит худи поверх футболки и куртку.
- **Страдает непереносимостью лактозы.**



Полиморфизм rs4988235

- Интрон гена MCM6 - влияет на экспрессию гена LCL
- Генотип (C;C) соответствует лактозной непереносимости (европ. популяция)

T > C gain-of-function

SNP name	motif	allele 1 / allele 2	P-value 1	P-value 2	Fold change	up/down	alignment
rs4988235	AP2C_HUMAN.H10MO.A (pcm, pwm) AP2C_HUMAN 	T/C	0.002087	4.2e-05	49.57	up	tggcaatacagataagataatgTAGTCCCTGGCCTCaaaggaactctcctc tggcaatacagataagataatgTAGCCCCTGGCCTCaaaggaactctcctc
rs4988235	AP2A_HUMAN.H10MO.C (pcm, pwm) AP2A_HUMAN 	T/C	0.003936	0.000123	31.96	up	tggcaatacagataagataatGTAGTCCCTGGCCTCaaaggaactctcctc tggcaatacagataagataatGTAGCCCCTGGCCTCaaaggaactctcctc
rs4988235	ID4_HUMAN.H10MO.D (pcm, pwm) ID4_HUMAN 	T/C	0.006731	0.000397	16.97	up	tggcaatacagataagataatgTAGTCCCTGGCctcaaaggaactctcctc tggcaatacagataagataatgTAGCCCCTGGCctcaaaggaactctcctc
rs4988235	HESX1_HUMAN.H10MO.D (pcm, pwm) HESX1_HUMAN 	T/C	0.005534	0.00036	15.37	up	tggcaatacagataagataATGTAGTCCCTGGCCTcaaaggaactctcctc tggcaatacagataagataATGTAGCCCCTGGCCTcaaaggaactctcctc
rs4988235	ZBTB6_HUMAN.H10MO.D (pcm, pwm) ZBTB6_HUMAN 	T/C	0.000532	5.1e-05	10.4	up	tggcaatacagataAGATAATGTAGTCctggcctcaaaggaactctcctc tggcaatacagataAGATAATGTAGCCctggcctcaaaggaactctcctc
rs4988235	INSM1_HUMAN.H10MO.C (pcm, pwm) INSM1_HUMAN 	T/C	0.003396	0.000343	9.9	up	tggcaatacagataagataatgTAGTCCCTGGCCtcaaaggaactctcctc tggcaatacagataagataatgTAGCCCCTGGCCtcaaaggaactctcctc

Полиморфизм rs12913832

Находится в интроне гена HERC2 - энхансере гена OCA2

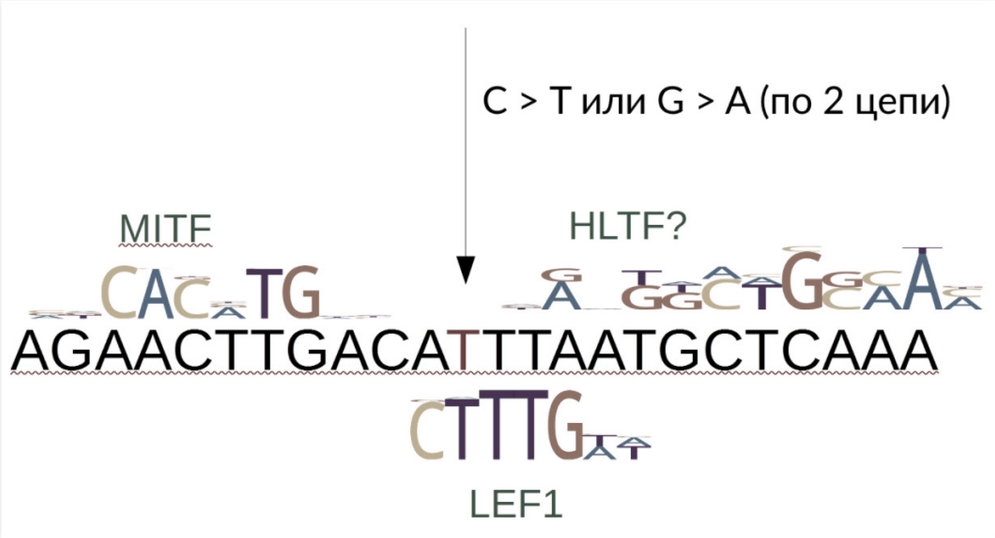
A:A карие глаза, 80%

A:G карие глаза

G:G голубые глаза, 99%

Mijke Visser *et al.*, *Genome Res.* 2012

Экспериментально показано дифференциальное связывание нескольких факторов транскрипции

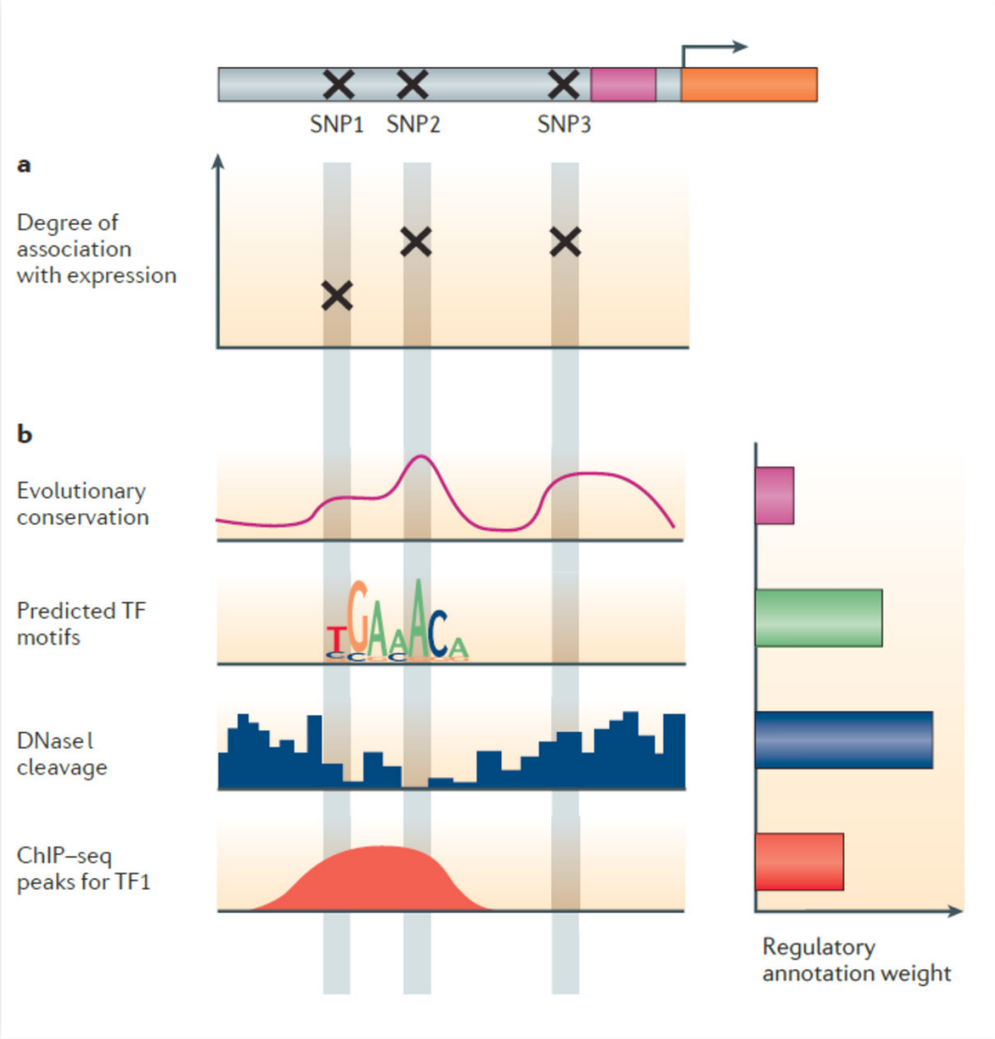


- но почему оно зависит от аллельного варианта?

G > A loss-of-function

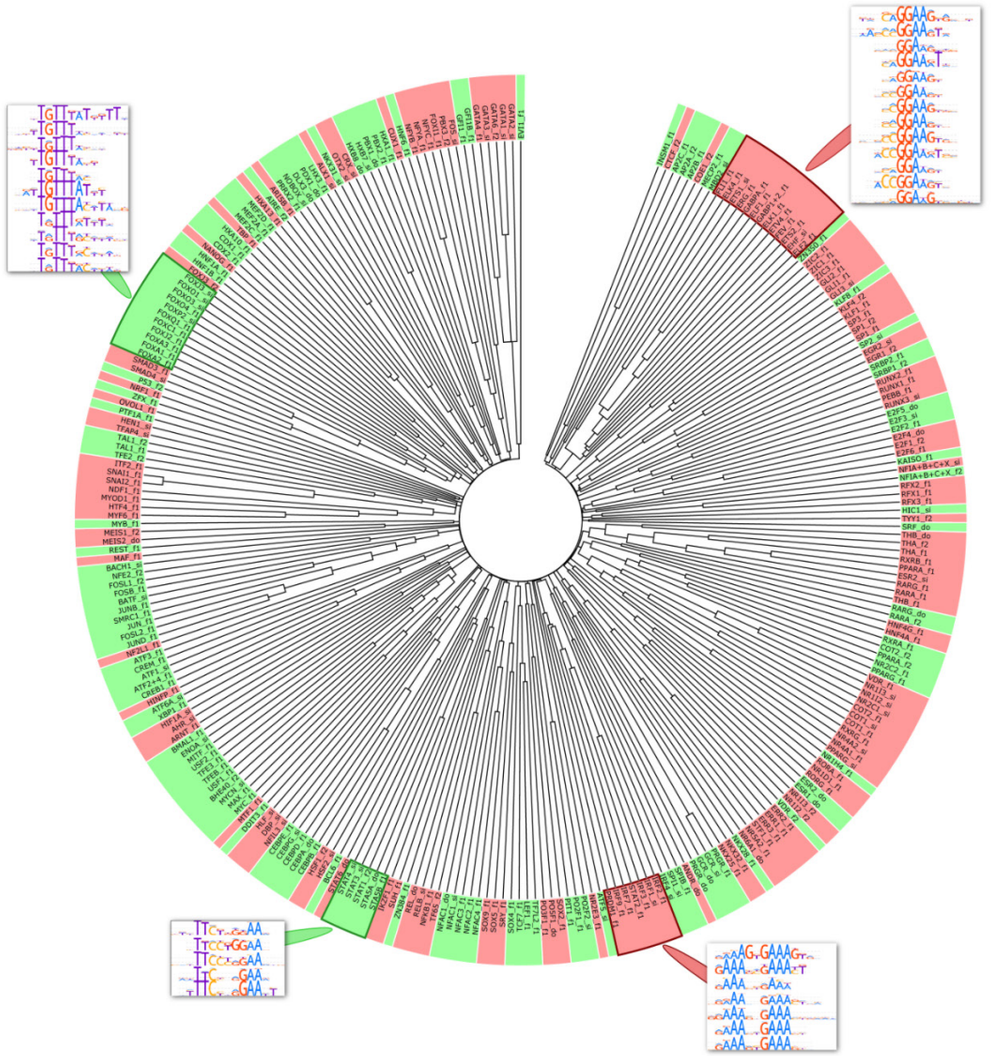
SNP name	motif	allele 1 / allele 2	P-value 1	P-value 2	Fold change	up/down	alignment
rs12913832	NKX25_HUMAN.H10MO.C (pcm, pwm) NKX25_HUMAN 	G/A	8.6e-05	0.00354	41.01	down	gaggccagtttcatttgagcaTTAAGTGtcaagttctgcacgctatcatca gaggccagtttcatttgagcaTTAAATGtcaagttctgcacgctatcatca
rs12913832	TBX5_HUMAN.H10MO.D (pcm, pwm) TBX5_HUMAN 	G/A	0.000129	0.004332	33.46	down	gaggccagtttcatttgagcattAAGTGTCAagttctgcacgctatcatca gaggccagtttcatttgagcattAAATGTCAagttctgcacgctatcatca
rs12913832	NKX23_HUMAN.H10MO.D (pcm, pwm) NKX23_HUMAN 	G/A	0.000244	0.006729	27.6	down	gaggccagtttcatttgagcATTAAGTGTCAagttctgcacgctatcatca gaggccagtttcatttgagcATTAAATGTCAagttctgcacgctatcatca
rs12913832	EOMES_HUMAN.H10MO.D (pcm, pwm) EOMES_HUMAN 	G/A	7.9e-05	0.001717	21.65	down	gaggccagtttcatttgagcATTAAGTGTCAAgttctgcacgctatcatca gaggccagtttcatttgagcATTAAATGTCAAgttctgcacgctatcatca
rs12913832	NKX32_HUMAN.H10MO.C (pcm, pwm) NKX32_HUMAN 	G/A	0.000201	0.004336	21.62	down	gaggccagtttcatttgagCATTAAGTGTCaagttctgcacgctatcatca gaggccagtttcatttgagCATTAAATGTCaagttctgcacgctatcatca

Подробная функциональная аннотация генетических вариантов



Levo and Segal, 2014, *Nat Rev Genet*

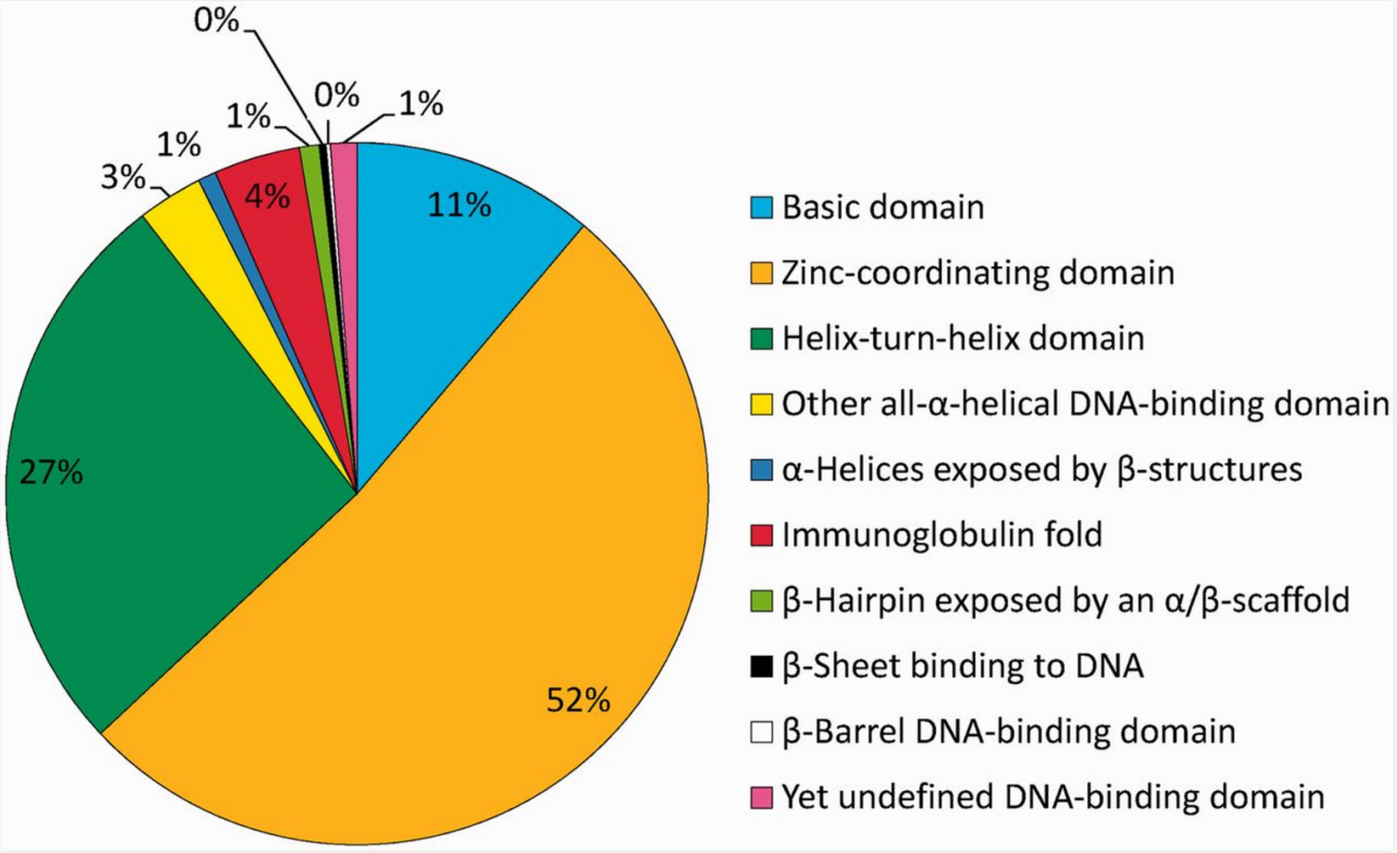
Сложность человеческого мотив-*ома* факторов транскрипции



Kulakovskiy, 2013

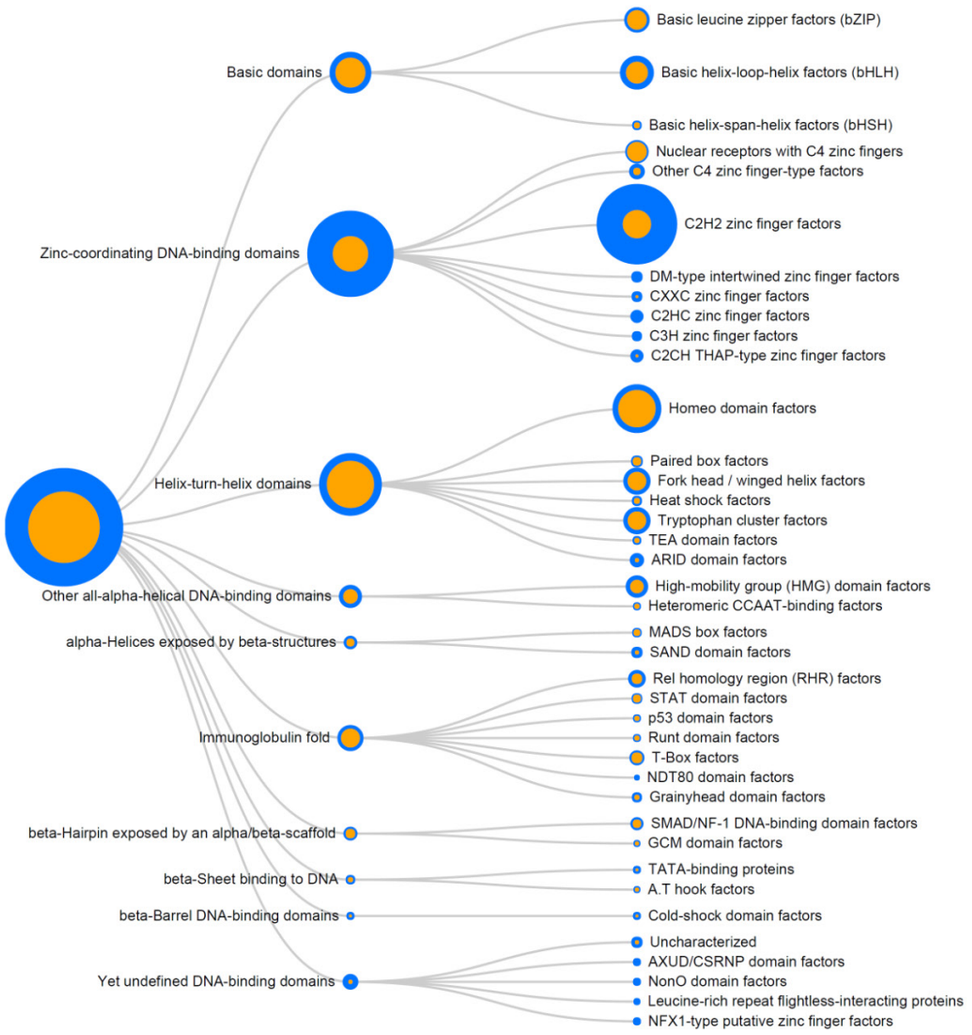
Насколько хорошо мы знаем "транскрипционфактор-*ом*"

1500-2000 из ~20000 белок-кодирующих генов кодируют факторы транскрипции



Wingender, 2012

Обновленный вид мотив-ома человека



Немного цифр

Мы знаем 600-800 мотивов для ~1500 факторов транскрипции человека, покрывая почти все структурные семейства

По оценкам систематического анализа мотивов в промоторах (FANTOM5) - мотив-ом известен не менее чем на 90%

Take-Home Message

Регуляторный код существует, но не единственен.

Мы практически полностью знаем мотив-*ом* человека, и горазо хуже знаем регуляцию.

Даже этих ограниченных знаний достаточно, чтобы анализ мотивов приносил практическую пользу.

Спасибо за внимание!

Инструменты для анализа мотивов живут здесь:

autosome.ru