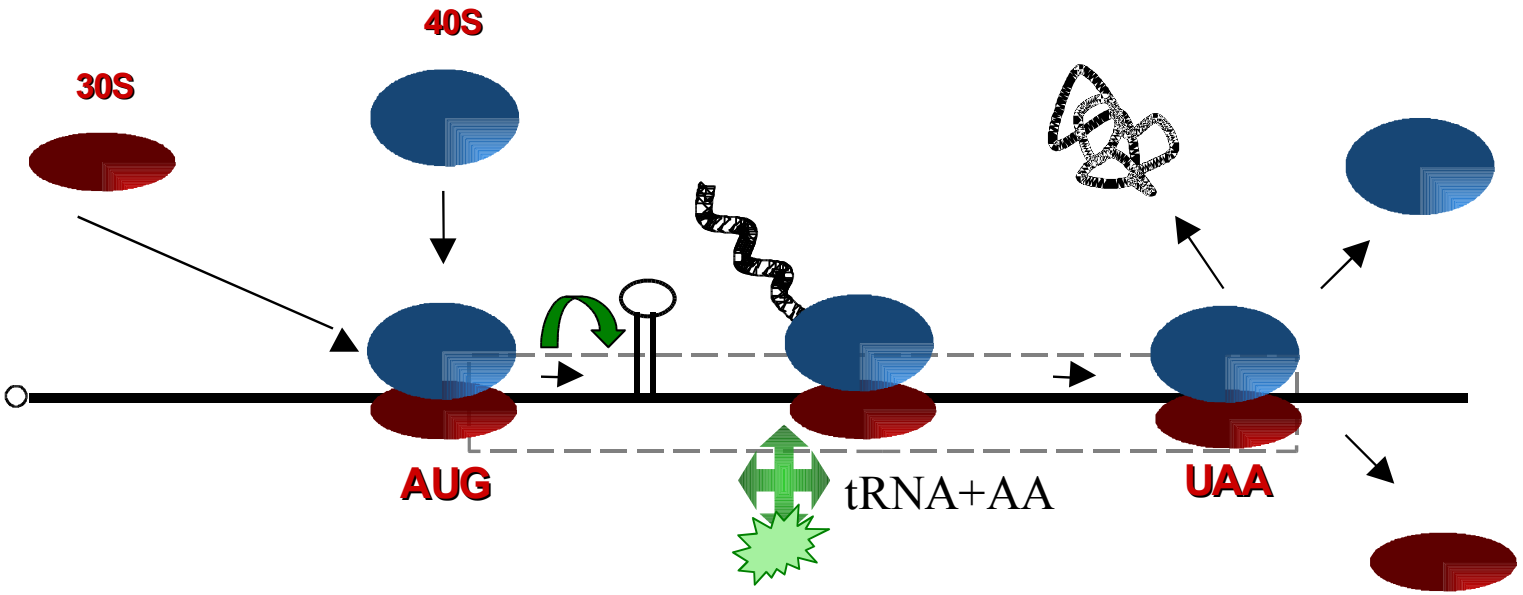
Теоретико-компьютерное исследование процесса трансляции

к.б.н. Матушкин Ю.Г.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

Схематическая модель трансляции прокариотической мРНК (мРНК, старт трансляции, шпилька, белок и стоп-кодон)



Универсальный генетический код

A	Ala	Аланин	GCA GCC GCG GCU
C	Cys	Цистеин	UGC UGU
D	Asp	Аспарагиновая кислота	GAC GAU
E	Glu	Глутаминовая кислота	GAA GAG
F	Phe	Фенилаланин	UUC UUU
G	Gly	Глицин	GGA GGC GGG GGU
H	His	Гистидин	CAC CAU
I	Ile	Изолейцин	AUA AUC AUU
K	Lys	Лизин	AAA AAG
L	Leu	Лейцин	UUA UUG CUA CUC CUG CUU
M	Met	Метионин	AUG
N	Asn	Аспарагин	AAC AAU
P	Pro	Пролин	CCA CCC CCG CCU
Q	Gln	Глутамин	CAA CAG
R	Arg	Аргинин	AGA AGG CGA CGC CGG CGU
S	Ser	Серин	AGC AGU UCA UCC UCG UCU
T	Thr	Треонин	ACA ACC ACG ACU
V	Val	Валин	GUA GUC GUG GUU
W	Trp	Триптофан	UGG
Y	Tyr	Тирозин	UAC UAU

Проанализирована связь между эффективностью экспрессии генов и нуклеотидным составом всех белок-кодирующих последовательностей для 78 одноклеточных организмов.

Показано, что не для всех организмов учет только частотно-кодонных характеристик адекватно отражает эффективность экспрессии.

Сконструирована мера - индекс эффективности экспрессии - учитывающая информацию как о кодонных характеристиках, так и о степени локальной комплементарности мРНК.

Анализ показал, что 78 видов разделяются на 5 групп, в зависимости от того, какой процесс вносит наибольший вклад в скорость элонгации.

Индекс эффективности элонгации имеет 2 слагаемых:
первое зависит от частот использования кодонов;
второе от локальных шпилек на мРНК

Качество нуклеотидного состава конкретной мРНК (i-ой) оценивается значением индекса эффективности элонгации **EEI(i),** который имеет **смысл среднего времени элонгации** всех учитываемых кодонов в гене:

$$EEI(i)=u_1T_a(i)+u_2T_e(i),$$

где u_1=0 или 1; u_2=0 или 1 – весовые коэффициенты, определяющие учет каждого слагаемого в значении индекса.

Время элонгации зависит от используемых кодонов:

частые/редкие <—>быстрые/медленные

Первое слагаемое T_a по кодонному составу гена оценивает среднее время, требуемое для размещения в А-сайте рибосомы изоакцепторной аминоацил-тРНК. Его значение подсчитывается по формуле

$$T_a(i)=\sum_{j=1}^{n_i}\beta_{\delta(i,j)}\Big/ n_i\,,\quad \beta_\delta=\frac{\sum_{m=1}^{C}\sqrt{\alpha_m}}{\sqrt{\alpha_\delta}}\,,$$

где величина $1/\beta_{\delta(i,j)}$ интерпретируется как оптимальная относительная концентрация аминоацил-тРНК, комплементарной j-ому учитываемому кодону, а $\alpha_{\delta(i,j)}$ и α_m имеют смысл частот использования кодонов $\delta(i,j)$ и m в некоторой выделенной подвыборке мРНК

Время элонгации зависит от количества и качества локальных шпилек, которые встречаются рибосоме вдоль мРНК

Второе слагаемое $T_e(i)$ оценивает по уровню автокомплементарности i-й мРНК среднее время, затрачиваемое рибосомой на стадию транслокации: $T_e(i)= t_{min}\cdot(1-p(i))+t_{max}\cdot p(i)$,

t_{min} - минимальное условное время транслокации,

t_{max} – максимальное условное время транслокации,

p(i) – вероятность реализации максимального условного времени транслокации, которую вычисляли по формуле

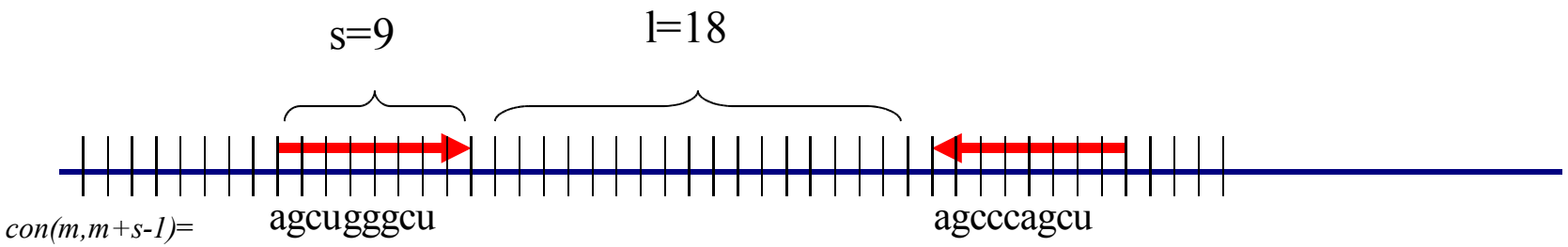
$$p(i) = \int_0^{LCI(i)} \frac{k^{n+1}x^n}{G(n+1)} e^{-kx} dx$$

$k=m/\sigma^2$, $n=(m/\sigma)^2$, где m и σ^2 , соответственно, математическое ожидание и дисперсия положительной случайной величины имеющей плотность распределения , G(n+1) – Gamma-функция,

LCI(i) - индекс локальной комплементарности.

Индекс локальной комплементарности есть мера числа локальных шпилек и их энергопотенциала

$$LCI(i,j) = \sum_{\substack{m:\quad m\le j\le m+s-1 \quad or \\ m+s+l-1\le j\le 2m+2s+l-2}} 10\left\{\sum_{s=s_{\min}}^{s_{\max}}\left[\sum_{l=l_{\min}}^{l_{\max}} -\psi\left(con(m,\ m+s-1), \overline{con(m+s+l-1, 2m+2s+l-2)}\right)\right]\right\}$$



-10ψ=17+34+17+18+29+29+34+17 – 67=128

aa, au, ua, ca, cu, ga, gu, cg, gc, gg
uu, au, ua, ug, ag, uc, ac, cg, gc, cc
9 9 11 18 17 23 21 20 34 29

Destabilasing weights of loops

-1000 -1000 -74 -59 -44 -43 -41 -41 -42 -43
-46 -49 -53 -56 -59 -61 -64 -67 -69 -71
-73 -75 -77 -79 -81 -83 -85 -87 -88 -89
-90 -91 -92 -93 -94 -95 -96 -97 -98 -99
-99 -99 -100 -100 -100 -100 -101 -101 -101 -101

D.N.Turner, N.Sugimoto (1988) Ann.Rev. Biophys. Biophys.Chem.17, 167-192

Индекс локальной комплементарности может подсчитываться без учета энергии вторичных структур

ζ-форма LCI. На участке длиной m_i нуклеотидов, подсчитывается усредненное количество комплементарных участков без учета энергии образования вторичных структур (в настоящей работе m_i равнялось утроенному количеству кодонов входящих в i-й ген плюс 53 нуклеотида со стороны его 3'- конца):

$$LCI\ \ \varsigma(i) = \frac{\sum_{m=1}^{m_i - s_{max} - l_{max}} \left\{ \sum_{s=s_{min}}^{s_{max}} \left[\sum_{l=l_{min}}^{l_{max}} \zeta\left(con\ (m, m+s-1), \overline{con\ (m+s+l-1, 2m+2s+l-2)} \right) \right] \right\}}{m_i - s_{max} - l_{max}},$$

где *con(i,j)* - контекст гена с i-го по j-й нуклеотиды и $\overline{con(i,j)}$ - комплементарный контекст гена с j-го по i-й нуклеотиды ($i \le j$), $\zeta(conext1, conext2)=1$, если слова conext1 и conext2 идентичны, в противном случае $\zeta(conext1, conext2)=0$. Длина учитываемого инвертированного повтора не меньше s_{min} и не больше s_{max}, расстояние между учитываемыми инвертированными повторами не меньше l_{min} и не больше l_{max} (в работе приняты $s_{min}=s_{max}=3$, $l_{min}=3$, $l_{max}=50$).

Индекс локальной комплементарности может подсчитываться с учетом энергии вторичных структур

ψ-форма LCI Подсчитывается усредненная энергия образования вторичных структур на участке длиной m_i нуклеотидов (в работе приняты $s_{min}=3$, $s_{max}=6$, $l_{min}=3$, $l_{max}=50$):

$$LCI\psi(i)=\frac{\sum_{m=1}^{m_i-s_{max}-l_{max}}\left\{\sum_{s=s_{min}}^{s_{max}}\left[\sum_{l=l_{min}}^{l_{max}}\psi\left(con(m,m+s-1),\overline{con(m+s+l-1,2m+2s+l-2)}\right)\right]\right\}}{m_i-s_{max}-l_{max}}$$

где ψ- энергия вторичной структуры, которая подсчитывается стандартным образом.

Время, затрачиваемое рибосомой на стадию транспептидации, полагалось равным для всех кодонов и всех генов и поэтому не учитывалось при построении EEI(i).

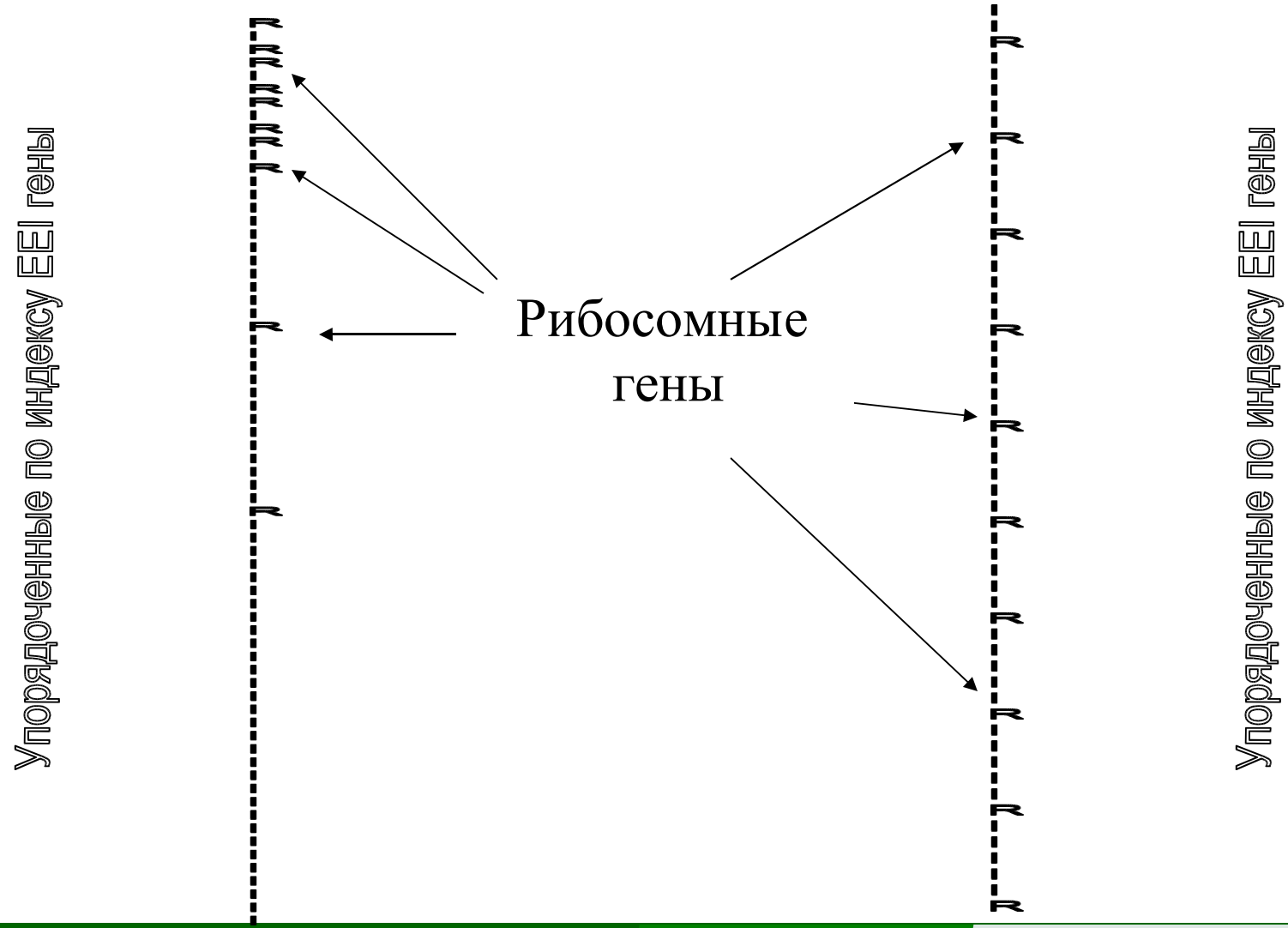
Критерий адекватности индекса элонгации EEI

Упорядочив Nгенов по увеличению значения EEI мы оцениваем его адекватность. Для этого мы подсчитываем вероятность того, что наблюдаемое распределение Kгенов рибосомных белков может быть получено по случайным причинам. Обозначим выборку рибосомных белков через $\Im$.Пусть n(g) обозначает номер гена g в полной выборке. Меру отклонения выборки $\Im$ от среднего положения расчитываем по формуле:

$$d(\Im) = 100 \cdot (2\sum_{g \in \Im} n(g) / K - N - 1) / (N - K)$$

Непосредственно проверяется, что $-100 \le d(\Im) \le 100$.Если выборка генов $\Im$ располагается случайным образом в установленном порядке генов, то можно подсчитать вероятность, $p=p(d \ge d(\Im))$,того, что произвольная выборка из K генов имеет отклонение не меньше$d(\Im)$.Если полученное значение pмало, то это интерпретируется как свидетельство неслучайного расположения рибосомных генов в фиксированном порядке всех генов. В качестве порога достоверности принято $\varepsilon=10^{-3}$.

Рибосомные гены считаются высокоэкспрессирующимися. Слева - индекс работает адекватно, справа - нет.



The shift from the center of the genes encoding ribosomal proteins in the total gene sample ordered according to values of various EEI modifications

Organism			Index of shift of genes encoding ribosomal proteins (GRP) from the center					
Eubacterium	Total number of genes	Number of GRPs	Genes ranked by EEI($u_1 = 1$, $u_2 = 0$) A	Genes ranked by EEI($u_1 = 0$, $u_2 = 1$) and LCIζ; ζ	Genes ranked by EEI($u_1 = 0$, $u_2 = 1$) and LCIψ; ψ	Genes ranked by EEI($u_1 = 1$, $u_2 = 1$) and LCIζ Aζ	Genes ranked by EEI($u_1 = 1$, $u_2 = 1$) and LCIψ Aψ	Stage that reflects maximally the expression efficiency;
A. tumefaciens Cereon	4556	44	74	11	33	25	53	A
B. halodurans	4066	55	78	-22	-33	-16	-32	A
B. subtilis	4178	53	85	0	2	0	15	A
B. melitensis	2059	51	70	0.1	15	38	40	A
C. muridarum	909	52	47	0	-19	49	0	A
C. pneumoniae	1052	55	63	-0.08	19	12	39	A
……another……	……	……	…16……	……	……	organisms	……	……
…								
E. coli O157 H7	5283	56	92	11	0.02	74	69	A
M. leprae	2720	53	70	-22	-27	45	47	A
Synechocystis PCC6803	3168	54	***47***	19	-35	44	-40	A/Aζ
V.cholerae	3828	59	95	0	2	11	17	A
B .burgdorferi	850	54	-19	49	-4	42	-30	ζ
Buchnera sp.ASP	569	54	-47	63	0	45	-25	ζ
C. jejuni	1654	53	-38	67	-39	55	-53	ζ
H. pylori J99	1491	53	-38	54	-18	48	-26	ζ
H. pylori 26695	1566	53	-36	50	-23	43	-31	ζ
M. genitalium G37	480	51	-62	56	-48	0	-62	ζ
M. pulmonis	782	56	-0.07	62	-0.07	53	-19	ζ
U.urealiticum	611	51	-15	77	-60	67	-58	ζ

The shift from the center of the genes encoding ribosomal proteins in the total gene sample ordered according to values of various EEI modifications

Organism			Index of shift of genes encoding ribosomal proteins (GRP) from the center					
Eubacterium	Total number of genes	Number of GRPs	Genes ranked by EEI($u_1 = 1$, $u_2 = 0$) A	Genes ranked by EEI($u_1 = 0$, $u_2 = 1$) and LCIζ; ζ	Genes ranked by EEI($u_1 = 0$, $u_2 = 1$) and LCIψ; ψ	Genes ranked by EEI($u_1 = 1$, $u_2 = 1$) and LCIζ Aζ	Genes ranked by EEI($u_1 = 1$, $u_2 = 1$) and LCIψ Aψ	Stage that reflects maximally the expression efficiency;
A. aeolicus	1522	55	32	63	-14	66	-10	Aζ
C. crescentus	3737	53	58	53	51	77	73	Aζ/Aψ
C. perfringens	2660	55	-57	78	-47	83	-57	Aζ
C. acetobutylicum	3672	59	50	45	50	74	34	Aζ
D. radiodurans	2936	54	64	49	42	76	52	Aζ
L. lactis subsp lactis	2263	56	80	53	-42	83	-63	Aζ/A
M. tuberculosis H37Rv	4187	57	32	17	0.08	51	21	Aζ
S. aureus Mu50	2714	55	75	44	-47	80	-46	Aζ
S. aureus N315	2593	55	75	45	-47	80	-46	Aζ
S. coelicolor A3(2)	7509	61	0	59	59	66	65	Aζ/Aψ
T.maritima	1846	51	44	62	24	69	37	Aζ
……another……	…….	……	…….	…..8……	……	organisms	……..	……
X. axonopodis pv citri str 306	4311	54	50	63	56	82	78	Aζ/Aψ
N. meningitidis MC58	1989	56	11	34	47	12	79	Aψ
N. meningitidis Z2491	2121	55	11	27	47	5	77	Aψ
S. typhimurium LT2	4451	57	84	26	31	43	86	Aψ/A
X._campestris	4181	54	26	65	64	79	79	Aψ/Aζ
X.fastidiosa	2766	55	-38	11	42	37	64	Aψ
P. aeruginosa PA01	5564	53	-75	83	84	84	85	Aψ/ψ/ζ/Aζ

The shift from the center of the genes encoding ribosomal proteins in the total gene sample ordered according to values of various EEI modifications

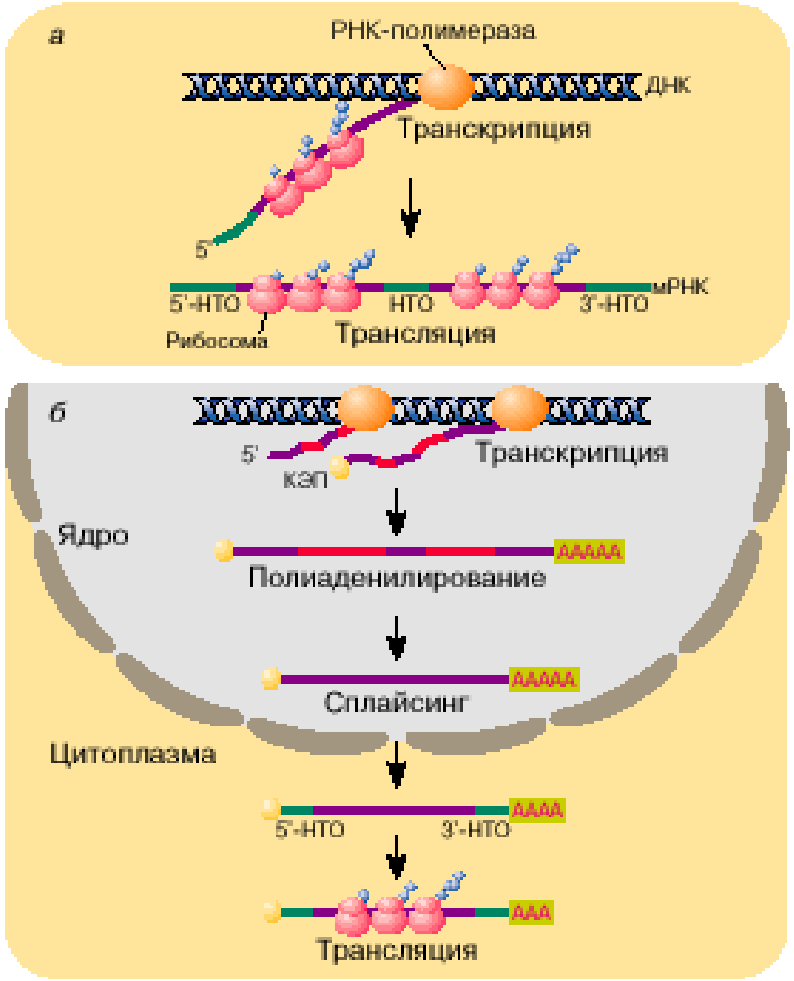
Organism	Total number of genes	Number of GRPs	Index of shift of genes encoding ribosomal proteins (GRP) from the center: Genes ranked by EEI($u_1 = 1, u_2 = 0$) A	Genes ranked by EEI($u_1 = 0, u_2 = 1$) and LCIζ; ζ	Genes ranked by EEI($u_1 = 0$, $u_2 = 1$) and LCIψ; ψ	Genes ranked by EEI($u_1 = 1, u_2 = 1$) and LCIζ Aζ	Genes ranked by EEI($u_1 = 1, u_2 = 1$) and LCIψ Aψ	Stage that reflects maximally the expression efficiency;
Arhaea								
A.fulgidus	2407	61	50	40	0	61	29	Aζ
A.pernix K1	2694	57	54	34	21	73	56	Aζ
Halobacterium	2058	56	-27	31	24	37	28	Aζ
M. mazei strain Goe1	3371	60	50	44	-29	63	-50	Aζ
M. jannaschii	1715	65	-6	68	-18	77	-20	Aζ
M. t-autotrophicum DeltaH	1869	61	26	56	25	66	27	Aζ
M. acetivorans str. C2A	4407	60	59	37	-35	62	-45	Aζ/A
M. kandleri AV19	1607	60	25	37	21	61	36	Aζ
P. aerophilum	2605	70	0	30	15	34	23	Aζ/ζ
P.abyssi	1760	59	45	44	-22	55	9	Aζ
P.horikosii OT3	2095	52	42	54	-5	69	14	Aζ
S. solfataricus	2977	65	37	45	12	59	19	Aζ
S. tokodaii	2826	65	35	41	16	56	21	Aζ
T. volcanium	1490	58	13	60	-13	61	-12	Aζ/ζ
T. tengcongensis	2588	56	-27	49	-23	46	-31	ζ/Aζ
T.acidophilum	1478	51	21	50	-8	49	0	ζ/Aζ
Eukaryote								
S. pombe	4681	114	94	42	-55	93	-66	A/Aζ
S. cerevisiae	6306	132	99	-4	-28	29	-23	A

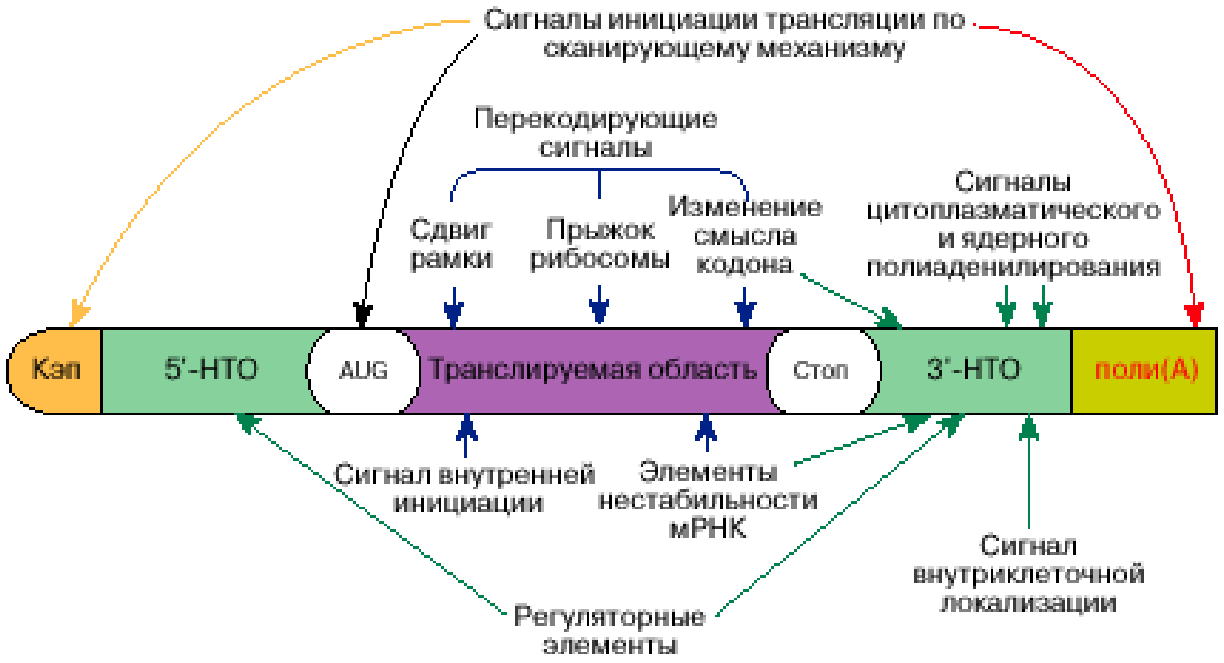
Выводы

Можно предположить, что в организмах ζ- ζ-А групп нал препятствия перед рибосомой запускает механизм, который каждый раз затрачивает определенную порцию ресурсов (временных, энергетических), удаляя с участка РНК определенной длины все помехи, независимо от их энергии и количества. В организмах из ψ- ψ-А групп, наоборот, механизм преодоления помех каким-то образом чувствует “ мощность” помехи и затрачивает на ее преодоление пропорциональное количество времени (и, возможно, энергии).

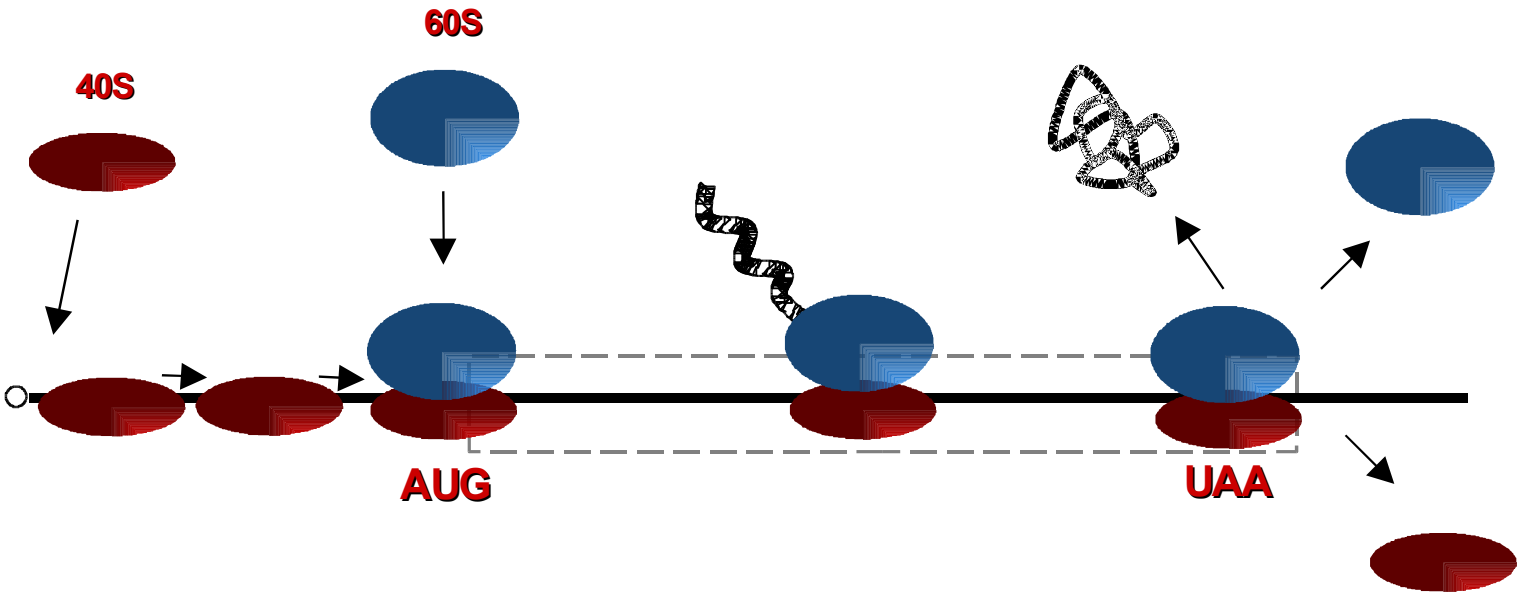
Выводы

Для объяснения потери в организмах из ζ- ψ- и групп, чувствительности стадии элонгации к кодонному составу гена, можно предложить следующие варианты объяснений: а) в данных организмах скорость размещения изоакцепторной аминоацил-тРНК в А-сайте рибосомы примерно одинаковая для всех кодонов, б) стадия размещения тРНК в А-сайте протекает параллельно с процессами, обеспечивающими нормальное протекание последующих этапов элонгации, и скорость первого процесса значительно быстрее второго, так что второй процесс как бы экранирует первый и тем самым обеспечивает эволюционную нейтральность кодонных мутаций.

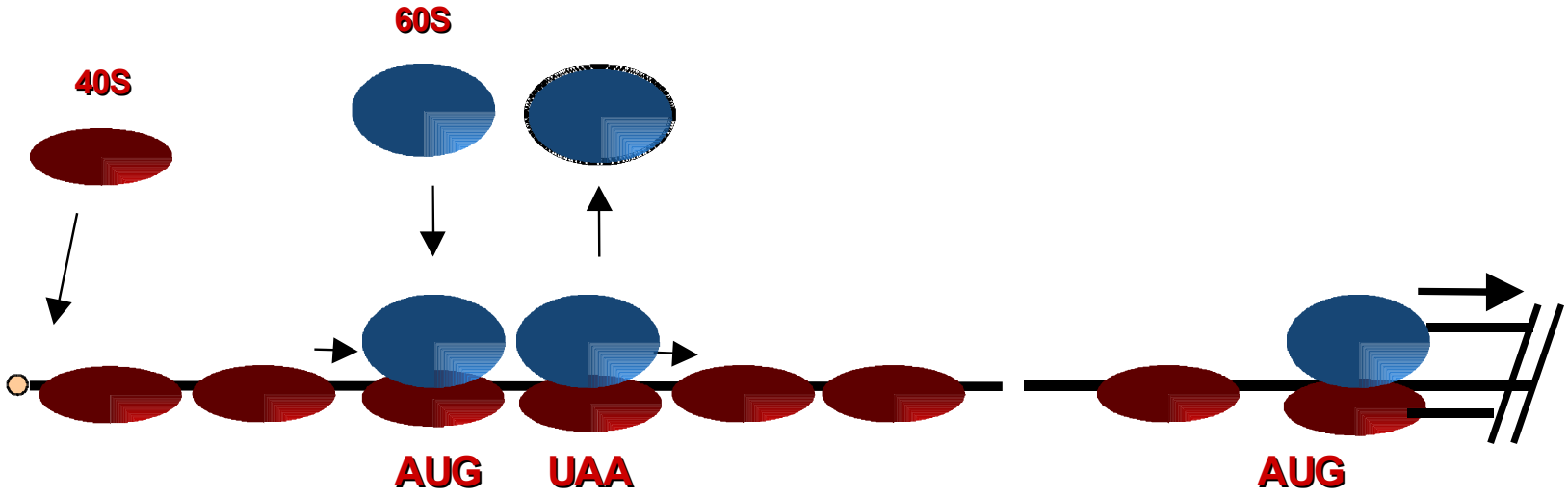

а
РНК-полимераза
ДНК
Транскрипция
5'
5'-НТО
НТО
3'-НТО
мРНК
Рибосома
Трансляция
б
5'
кэп
Транскрипция
Ядро
AAAAA
Полиаденилирование
AAAAA
Сплайсинг
Цитоплазма
AAAA
5'-НТО
3'-НТО
AAA
Трансляция


Сигналы инициации трансляции по сканирующему механизму
Перекодирующие сигналы
Сдвиг рамки
Прыжок рибосомы
Изменение смысла кодона
Сигналы цитоплазматического и ядерного полиаденилирования
Кап
5'-НТО
AUG
Транслируемая область
Стоп
3'-НТО
поли(А)
Сигнал внутренней инициации
Элементы нестабильности мРНК
Сигнал внутриклеточной локализации
Регуляторные элементы

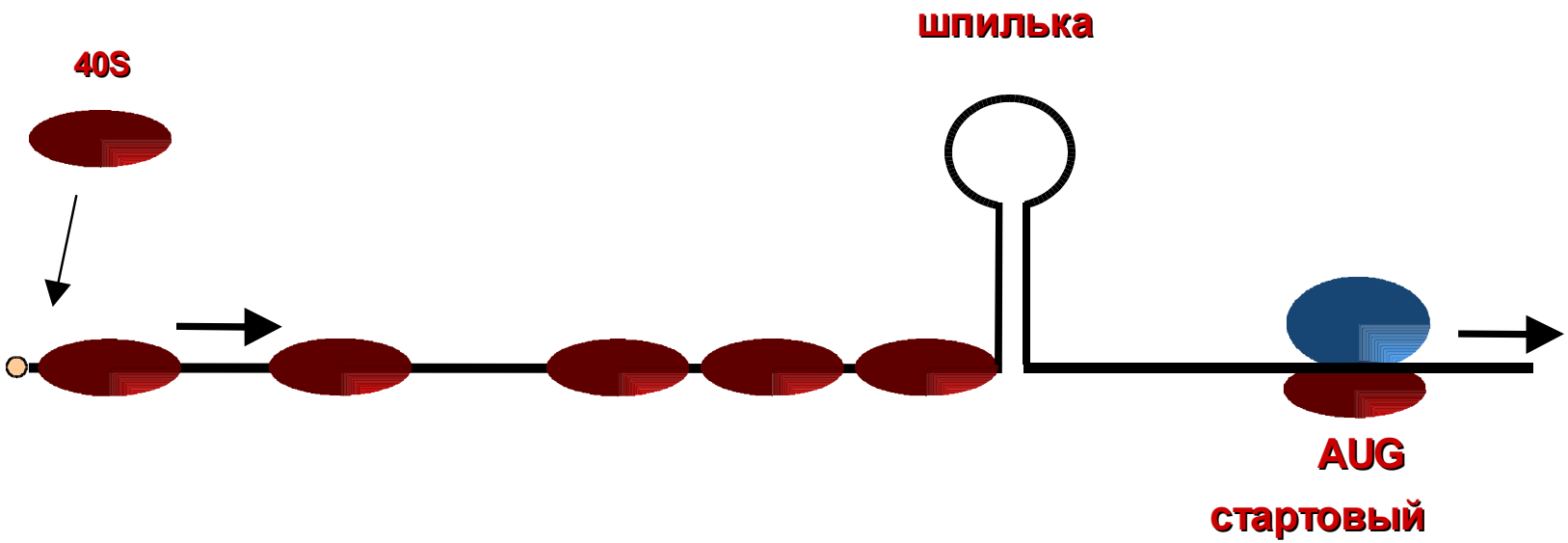
Сканирующая модель инициации трансляции эукариотических мРНК

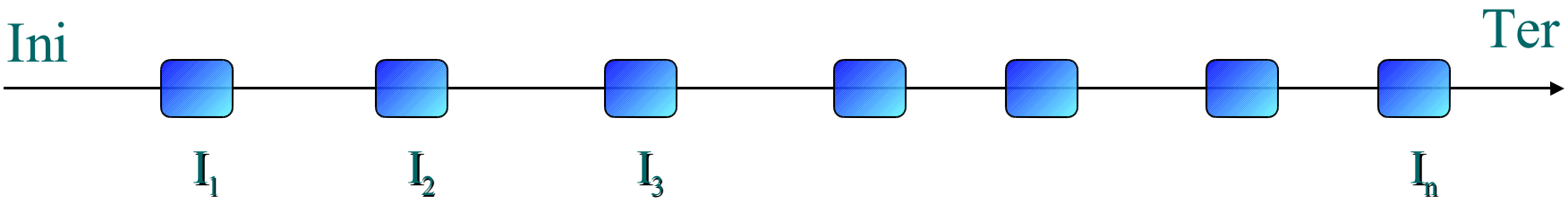


Эффективность трансляции эукариотических мРНК может быть снижена ложными стартами трансляции



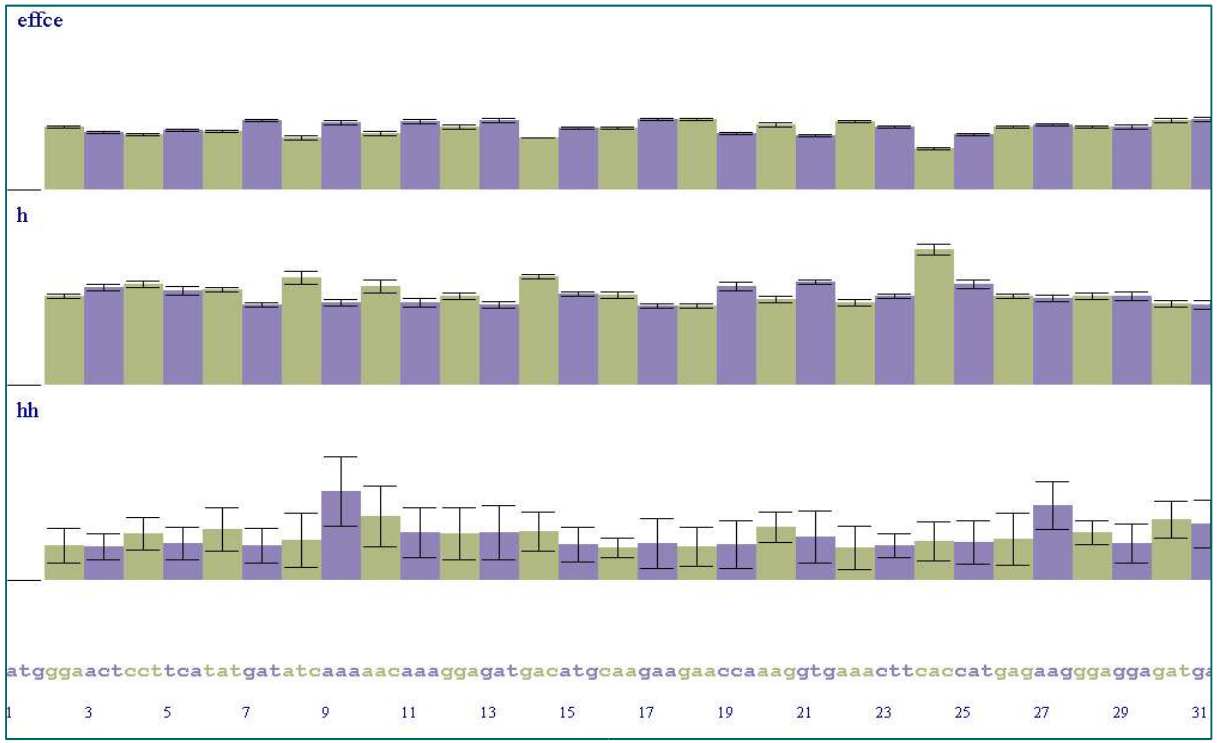
Эффективность трансляции эукариотических мРНК может быть снижена стабильными шпильками





В начальный момент каждый кодон получает оценку времени, которое проведет на нем рибосома. Далее, на каждом шаге расчета это время сравнивается с реально проведенным рибосомой на кодоне. Когда время превышает статистическую оценку - рибосома передвигается на следующий кодон. Инициация трансляции происходит, если свободен сайт инициации и время инициации меньше, чем время, оставшееся до перемещения одной из рибосом. Терминация - уход с последнего кодона.

Пример выдачи на экран результатов расчета модели трансляции

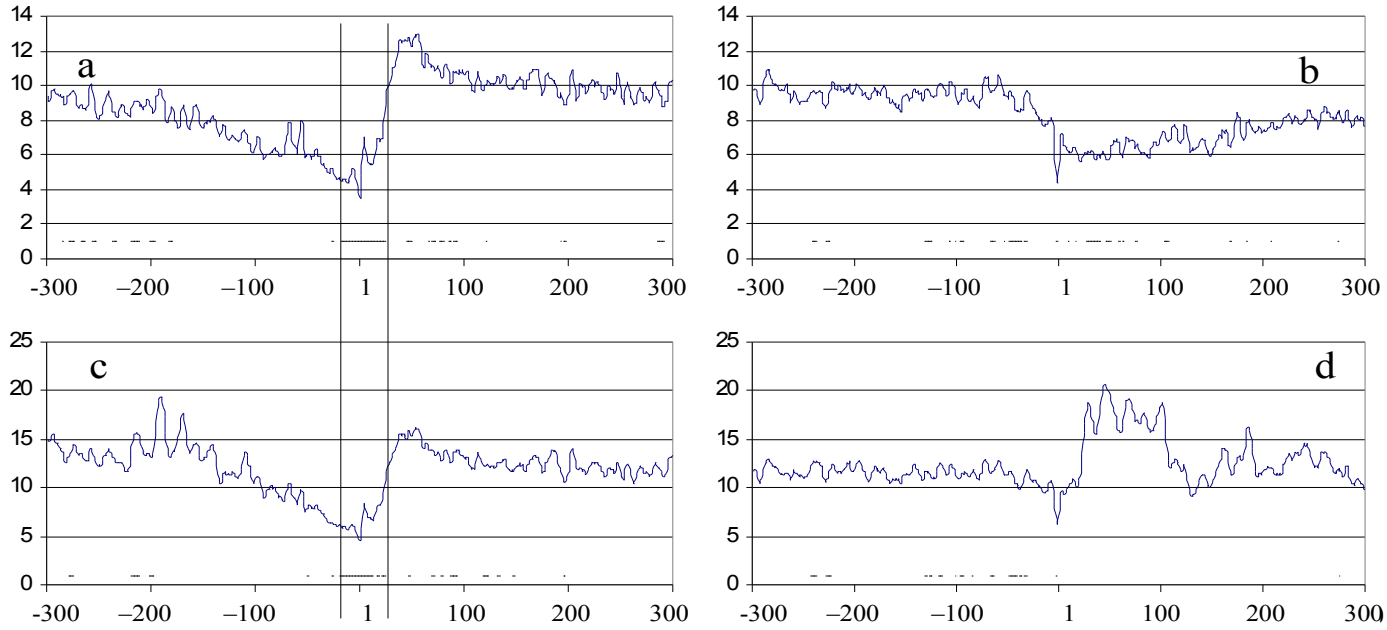


Effice - усредненная скорость прохождения рибосомой кодона; h - усредненная доля времени, которое кодон занят работающей рибосомой; hh - усредненная доля времени, которое рибосома проводит на кодоне из-за торможения стопкой рибосом впереди. Показана дисперсия средних значений.

Для *S. Cerevisiae* существует корреляция между уровнем экспрессии и стабильностью вторичных структур в области старта трансляции

Известно, что опознование сайта инициации трансляции зависит от контекста окружающей мРНК. Однако, многие эукариотические мРНК из баз данных содержат старт кодон AUG в неоптимальном контексте. Мы предположили, что свойства вторичной структуры мРНК вокруг стартового кодона также влияют на его опознание и на эффективность инициации трансляции.

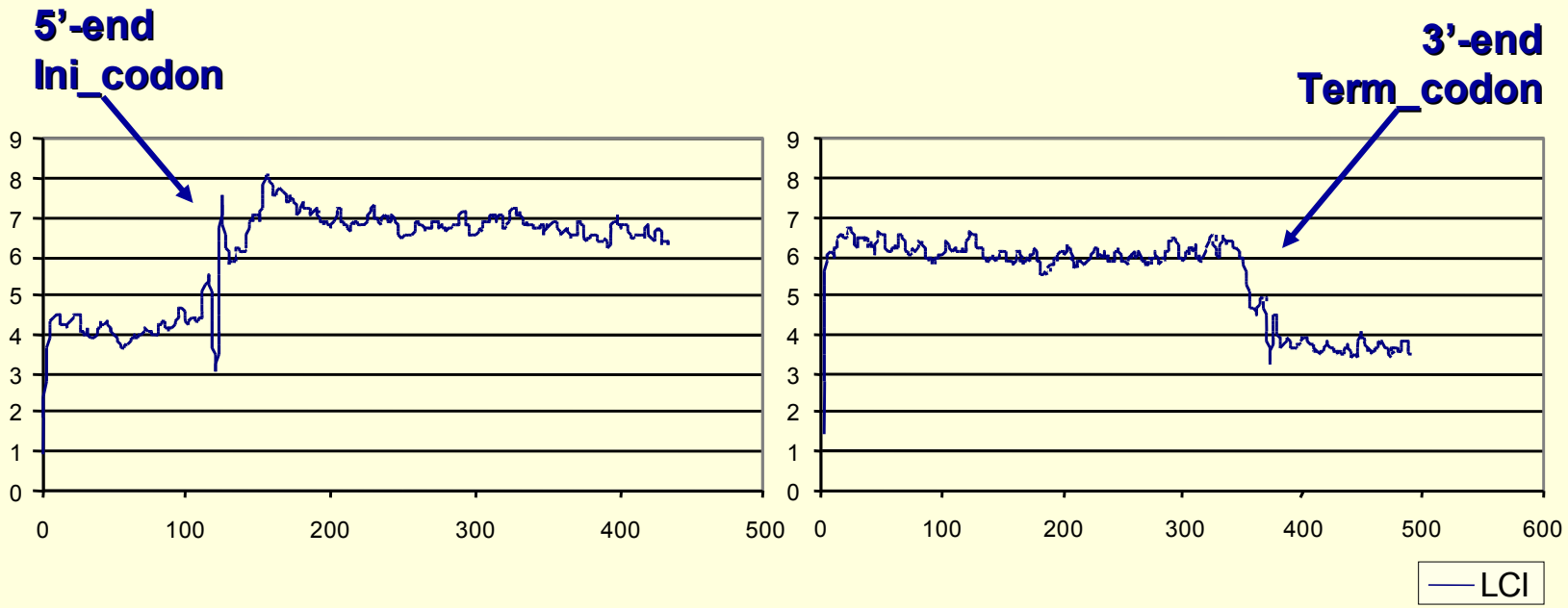
Оказалось, что стабильность предсказанной вторичной структуры в 5’ НТО снижена по сравнению с кодирующей частью. Кроме того, в этом районе обнаружилась достоверная отрицательная корреляция между значениями индекса эффективности экспрессии для кодирующей части и индекса локальной комплементарности для 5’ нетранслируемой области.



S. cerevisiae (6306 genes).

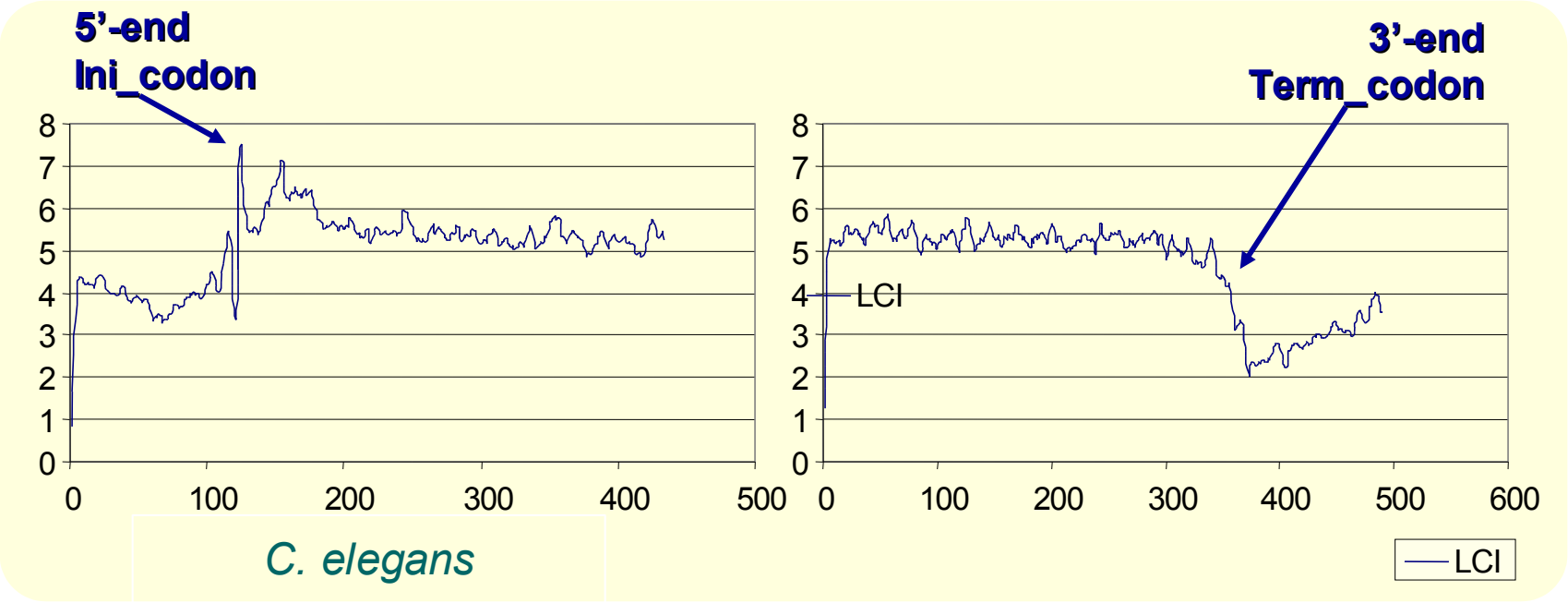
LCI-profils 5'- и 3'-областей. a,c – 5'-области, b,d -3'-области. По осям абсцисс отложены расстояния в нуклеотидах относительно инициаторного и терминаторного кодонов (+1,+3). По осям ординат отложены значения LCI-индексов. Черным пунктиром помечены позиции с достоверной корреляцией ($p<0.01$) LCI со значением EEI-индекса. a,b – расчеты произведены при учете совершенных шпилек, имеющих стемы в пределах 3-6, а петли в пределах 3-50 нуклеотидов, c,d - стемы в пределах 3-12, а петли в пределах 3-38 нуклеотидов.

LCI profile



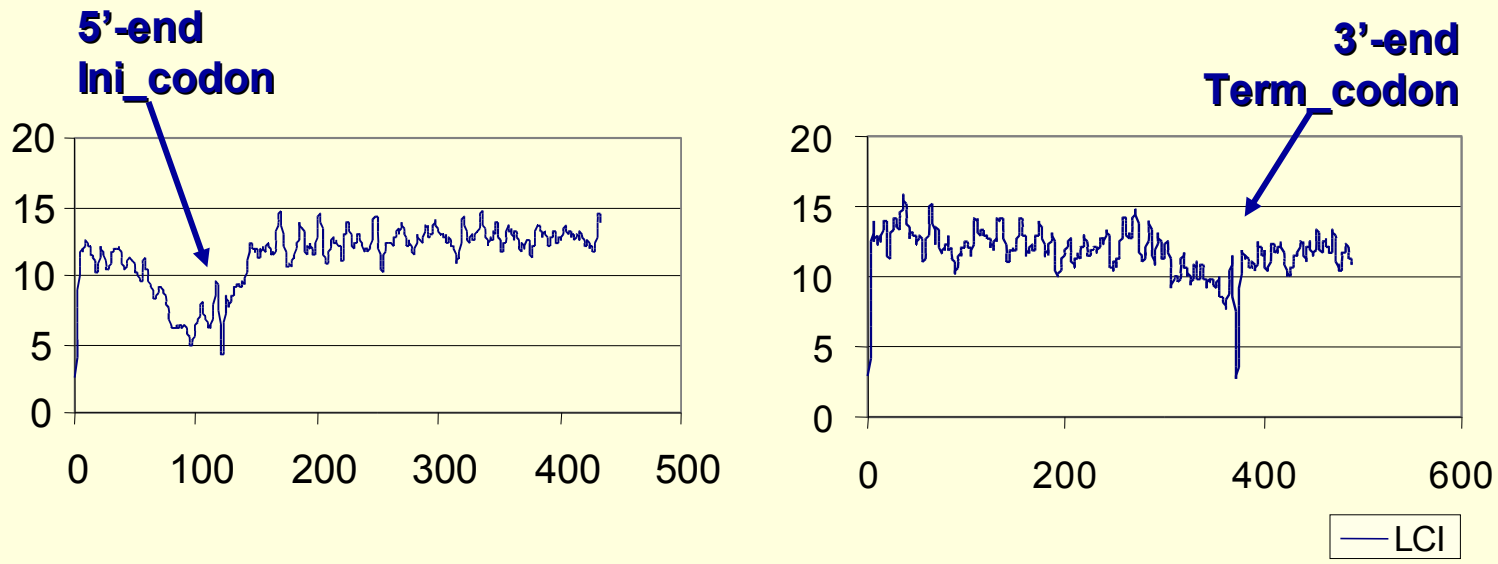
Arabidopsis (19700 genes)
Averaged out all the genes

LCI profile



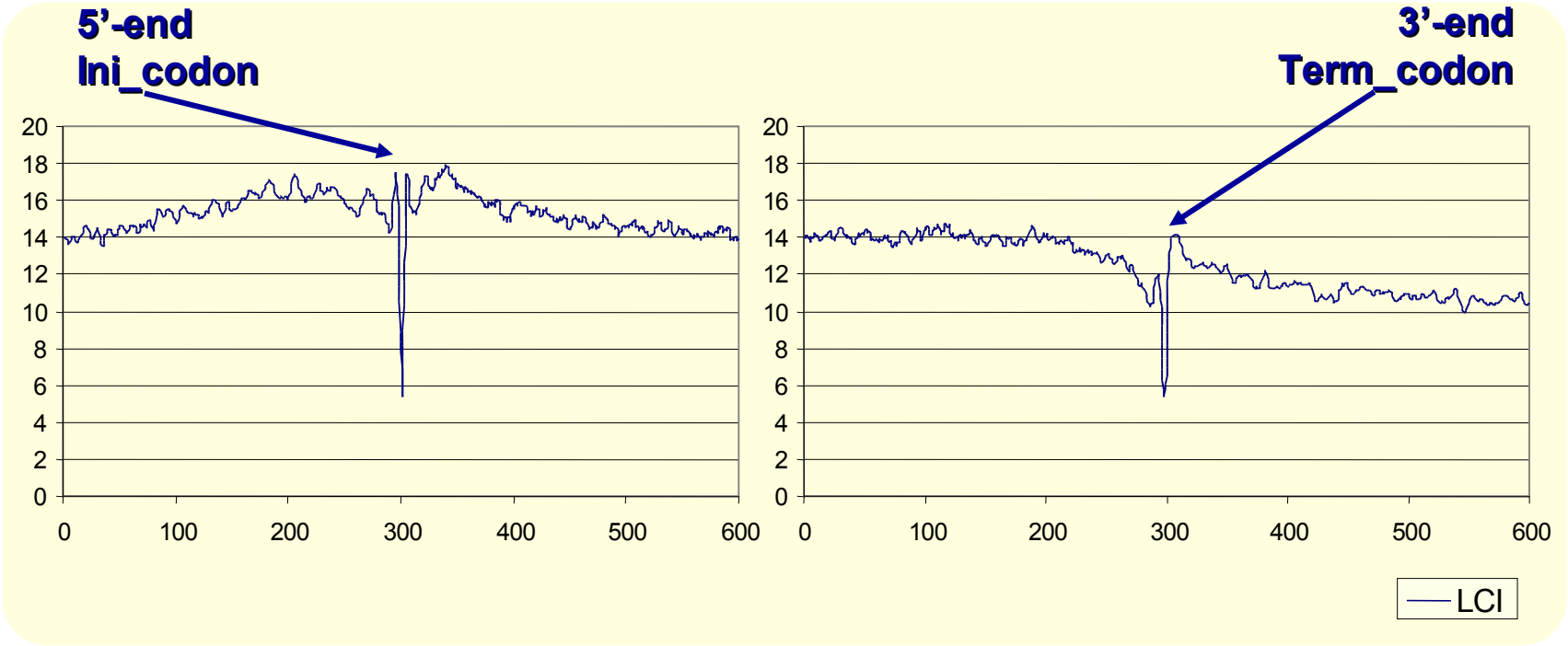
Averaged out the 17035 genes

LCI profile



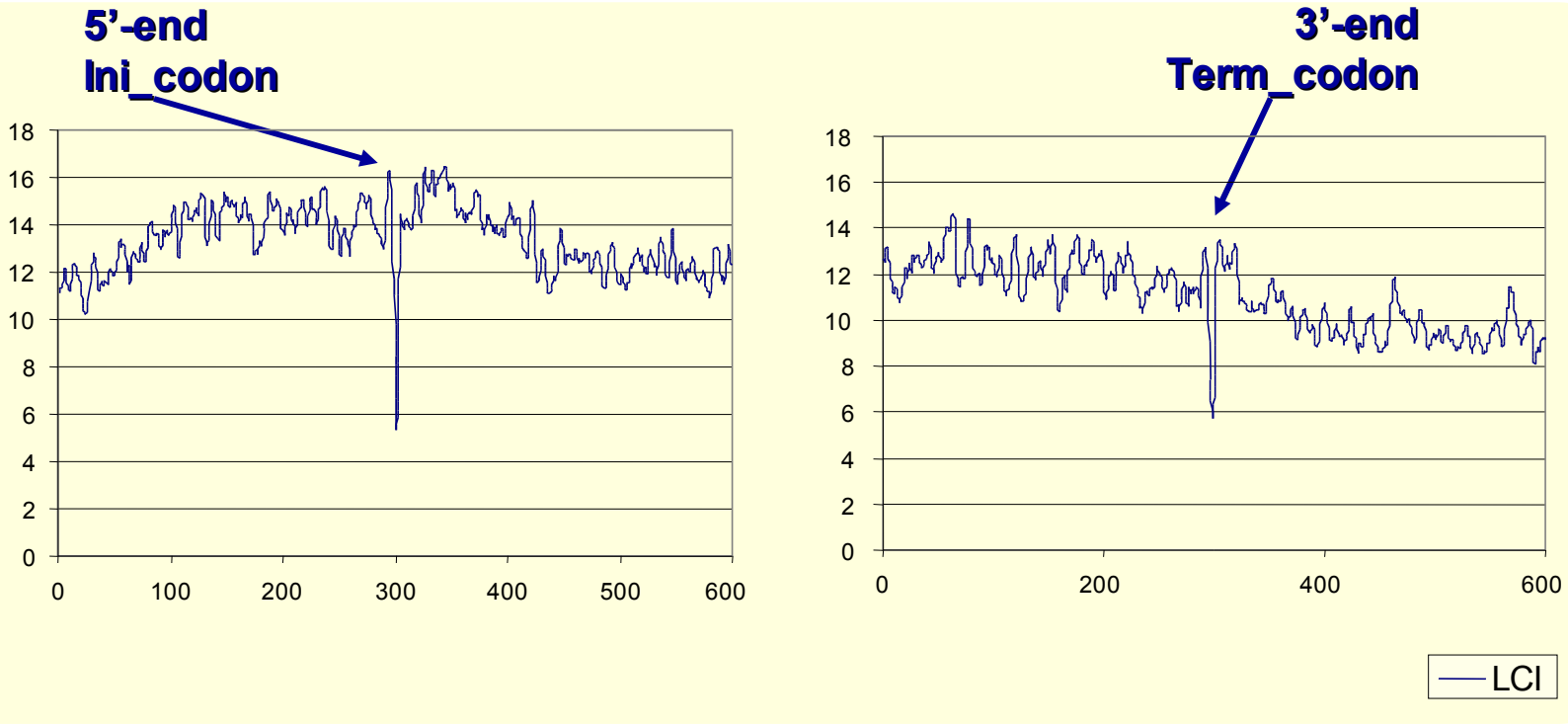
Archaea *Pyrobaculum aerophilum* (average out 2605 genes).

LCI profile



Homo sapiens
Averaged-out LCI profile of 34,485 genes

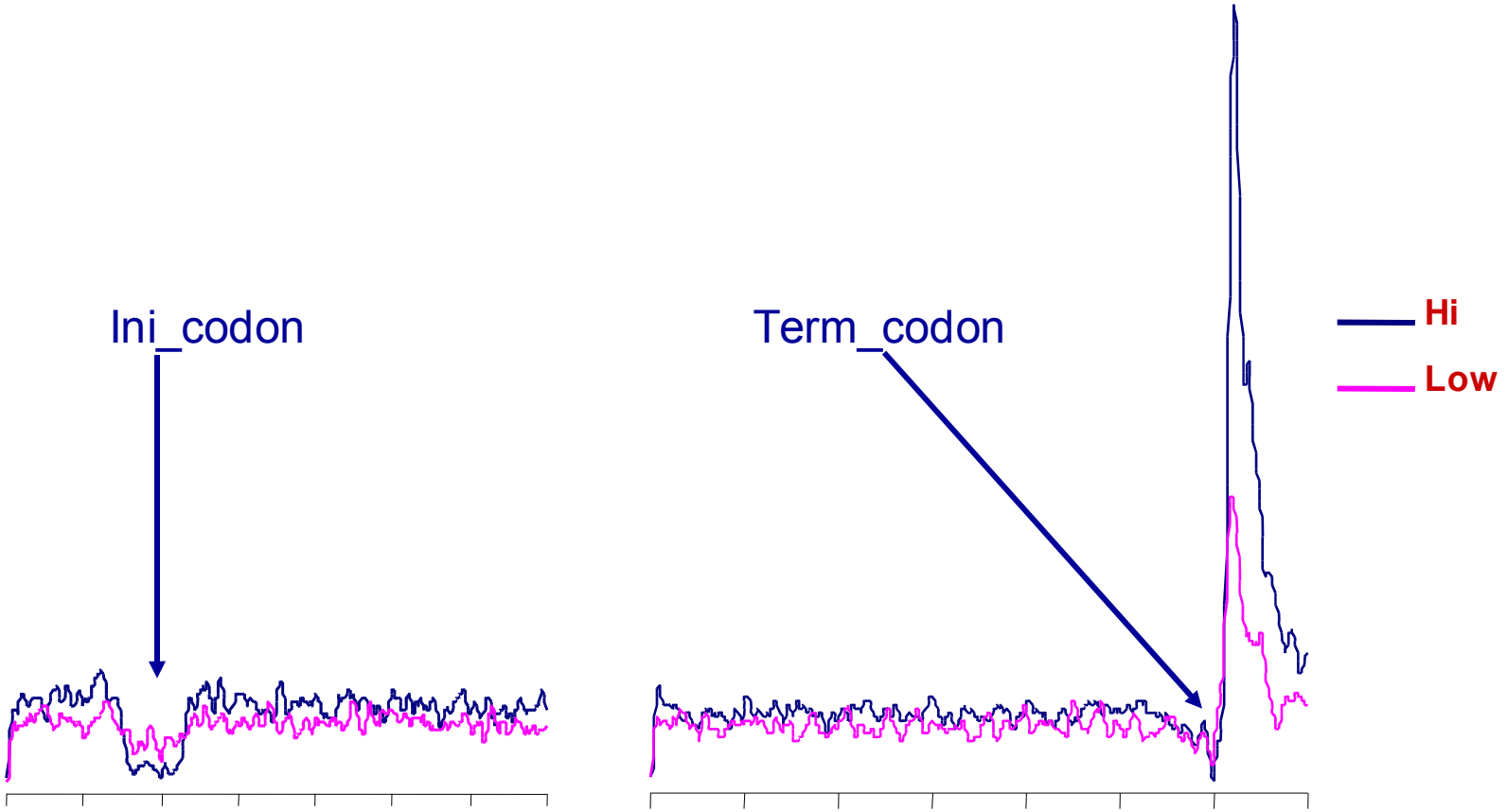
LCI profile



Mus Musculus

Averaged-out LCI profile of 3480 genes

E. coli K12 (4289 genes) LCI profile for Hi and Low expressed genes



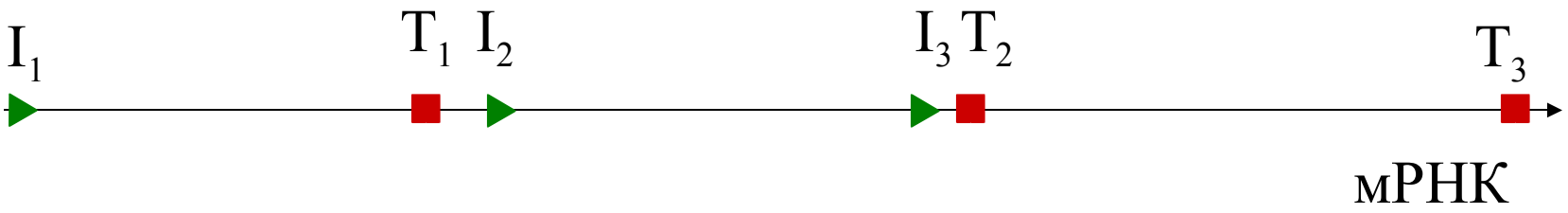
Выводы

Огромный пик на 3’-конце формируется реальными 3’-концами мРНК, т.к. на 3’-концах генов внутри оперонов он отсутствует. Возможно, это связано со стабильностью мРНК по отношению к нуклеазам или с усилением терминации трансляции и/или транскрипции.

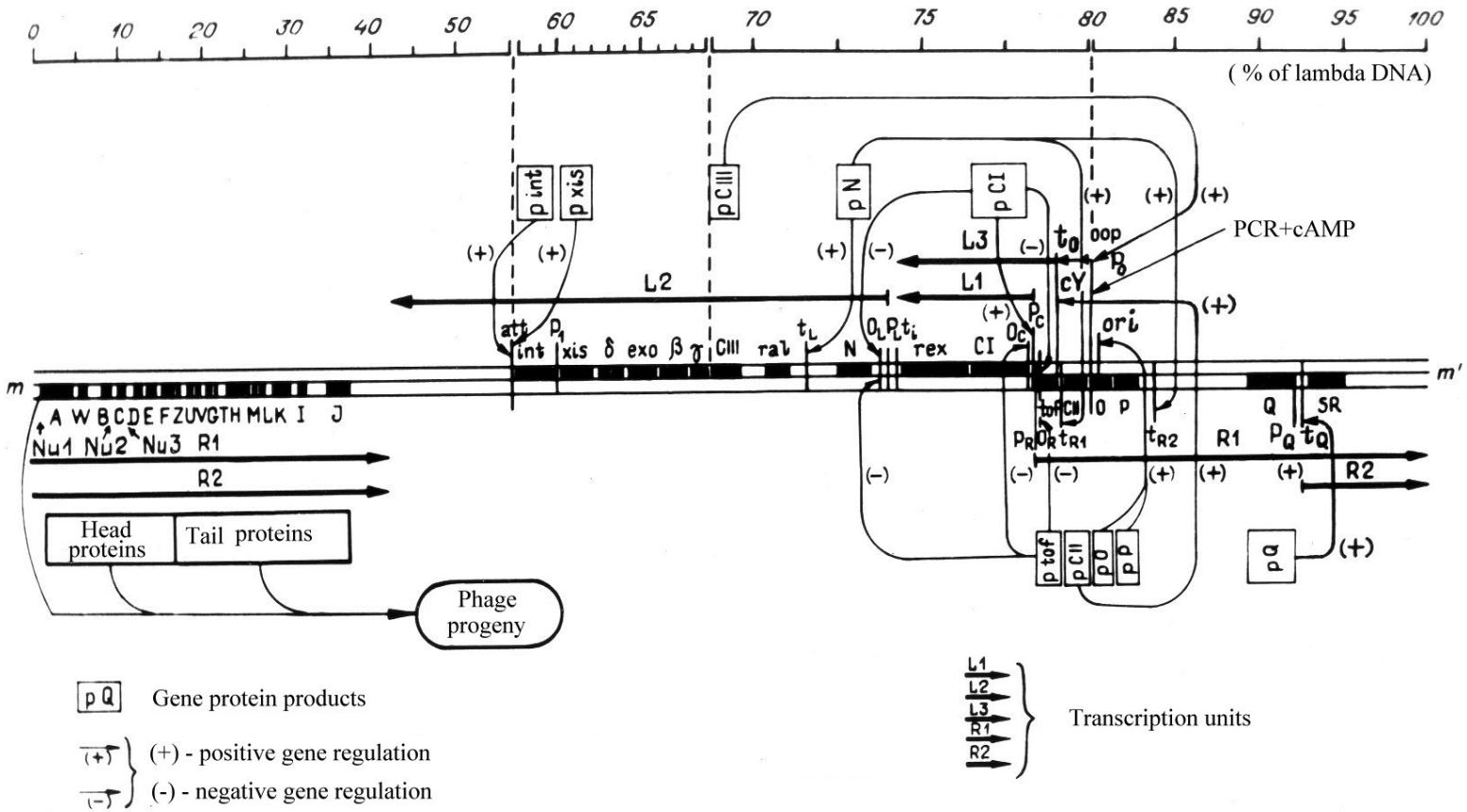
Яма на 5’-конце своими размерами коррелирует со стерическими размерами рибосомы. Также это связано с оперонной структурой генома E.coli, т.к. стартовый кодон одного гена может находиться в кодирующей части другого.

E. coli принадлежит к группе А, т.е. для адекватности индекса эффективности экспрессии в кодирующих районах нет нужды испольховать данные о локальной комплементарности. Однако, как можно видеть, вторичная структура может играть роль в изменении эффективности экспрессии гена, связанной с процессами инициации и терминации.

Оперонная структура

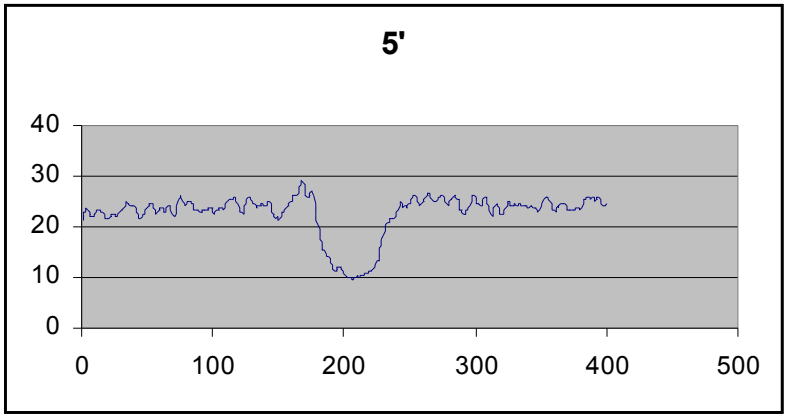


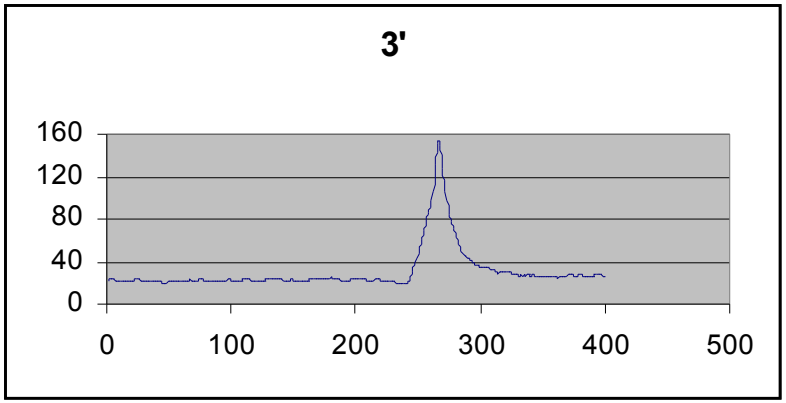
I - инициирующий кодон (ATG),
T - терминирующий кодон (TAA, TAG, TGA,)



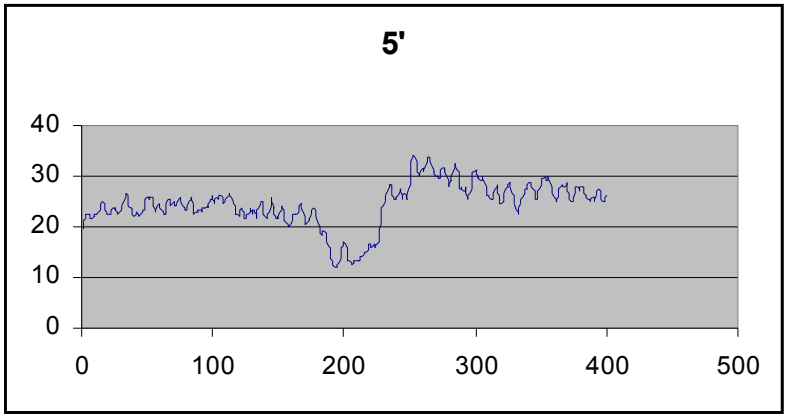
Genetic map of Lambda phage

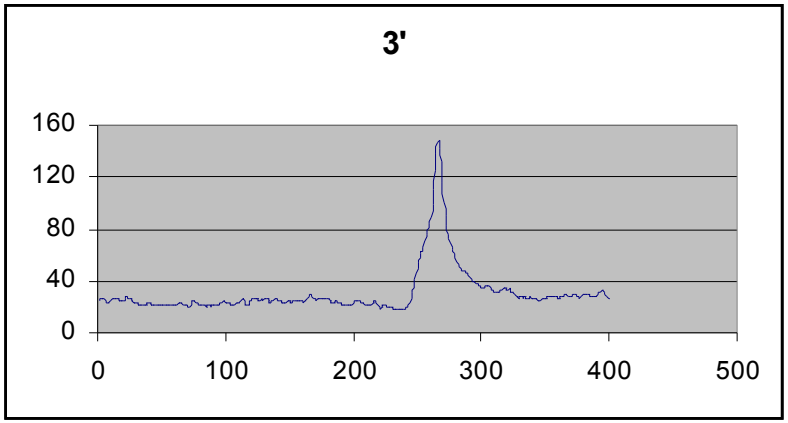
ОРС в конце оперона, рассматриваются все опероны *E.Coli*



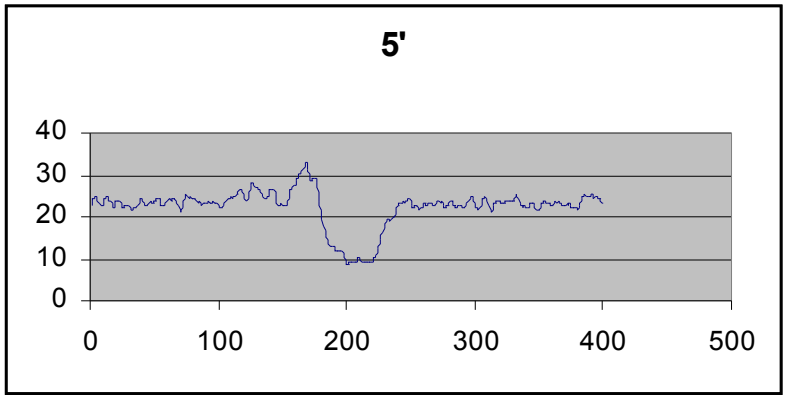


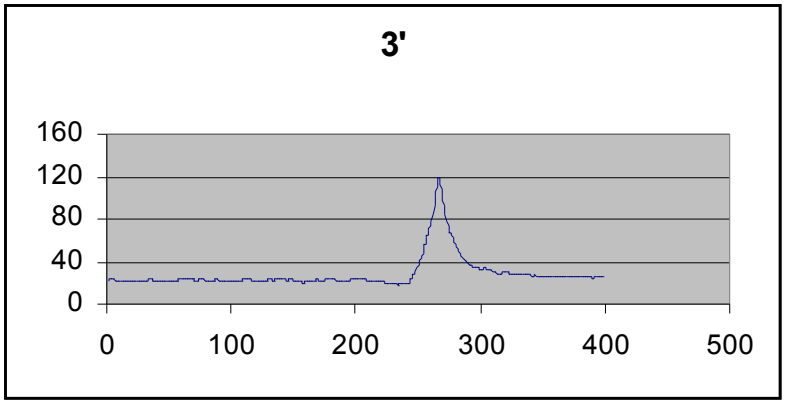
ОРС в конце оперона, в опероне больше одного гена



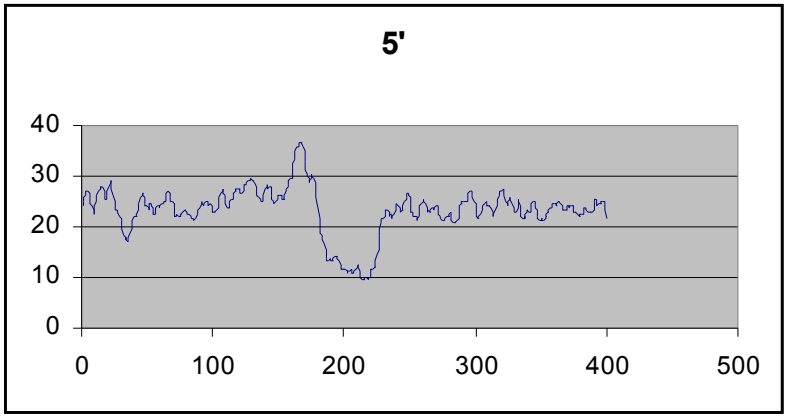


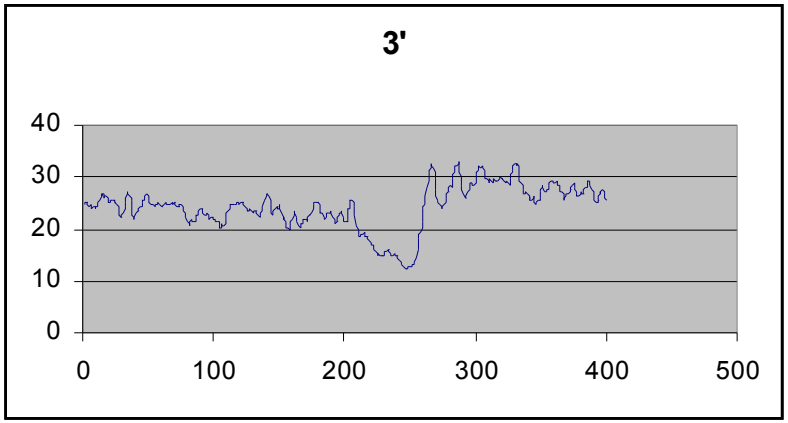
ОРС в начале оперона, рассматриваются все опероны *E.Coli*



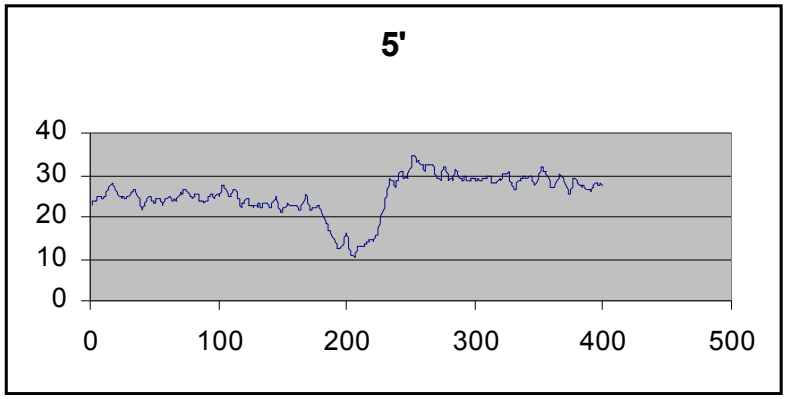


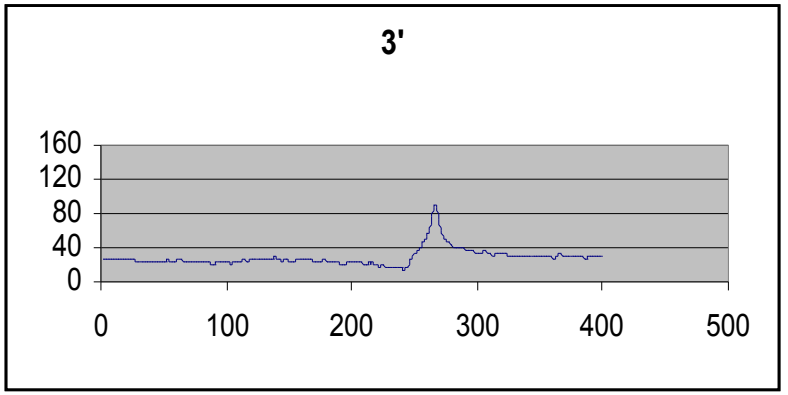
ОРС в начале оперона, в опероне больше одного гена



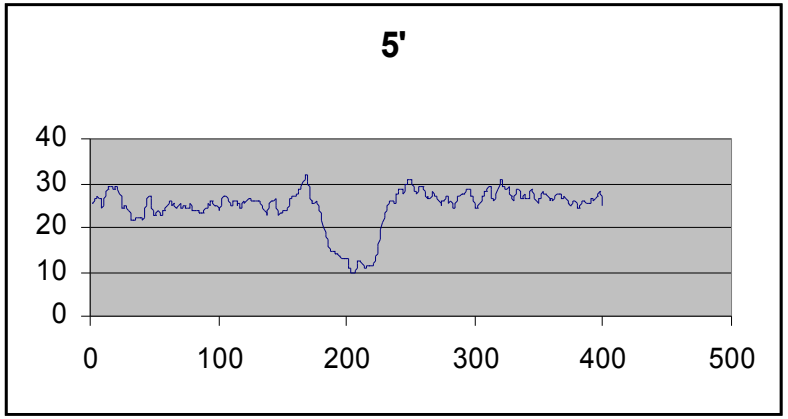


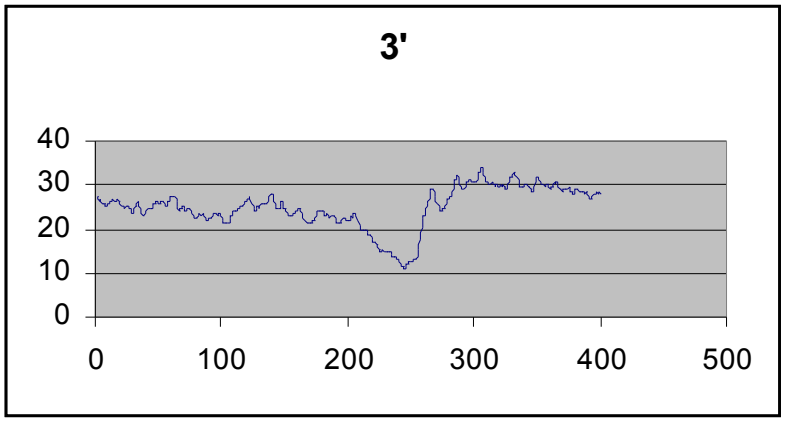
Не первая ОРС в опероне



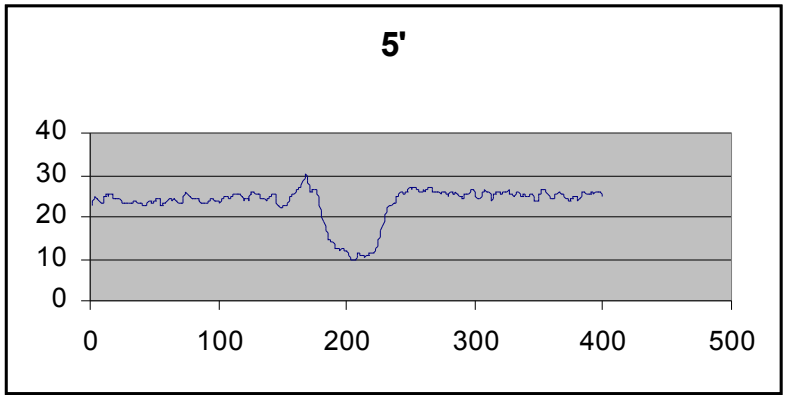


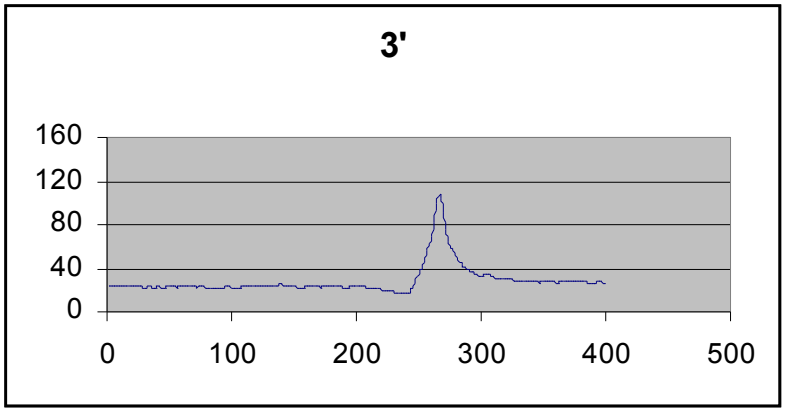
Не последняя ОРС в опероне



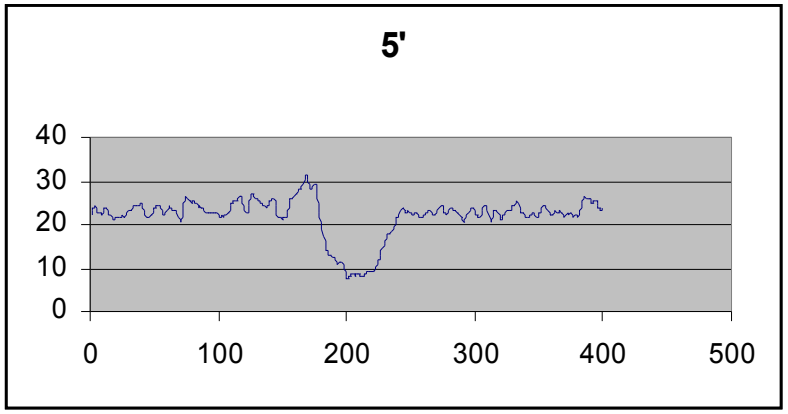


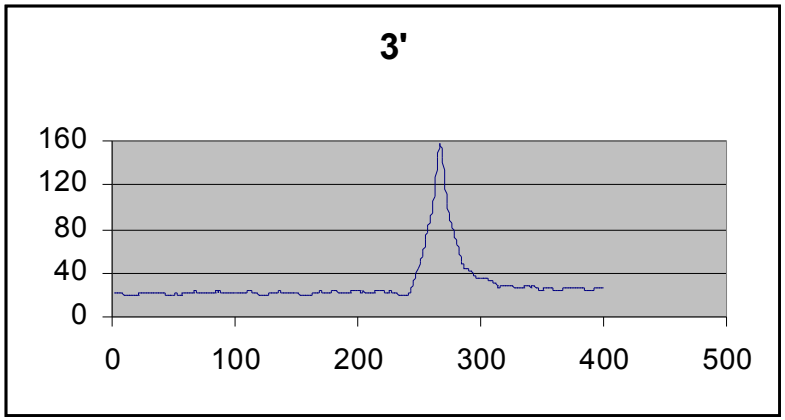
Рассматриваются все ОРС *E.Coli*



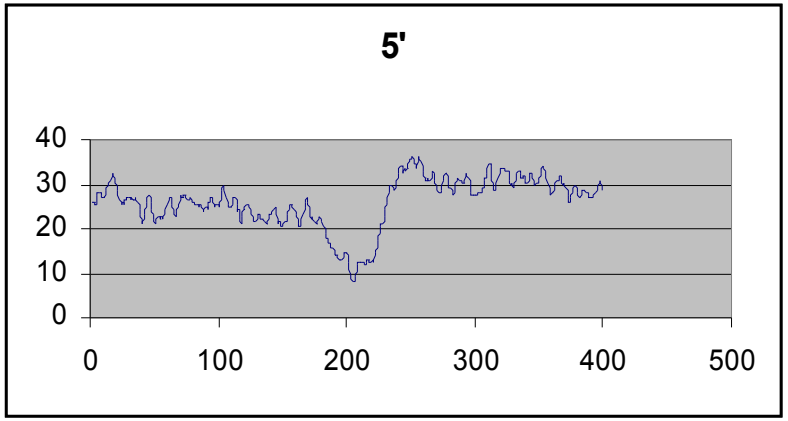


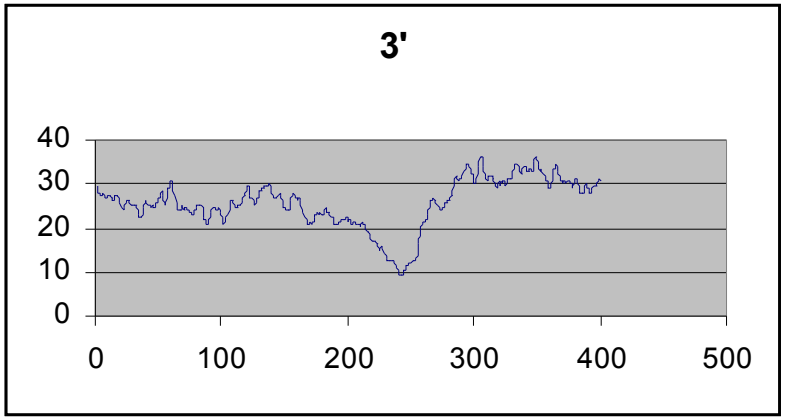
Одна ОРС в опероне





ОРС внутри оперона





Пересечение 2-х генов. 3’-конец левого лежит на 5’-конце правого, пересечение 20-10нук.