Институт проблем передачи информации РАН (http://www.iitp.ru)

Лаборатория математических методов и моделей в биоинформатике

(http://lab6.iitp.ru)

д.ф.-м.н. проф. зав лаб
Любецкий Василий Александрович
(lyubetsk@iitp.ru)

С 2000 года нами опубликовано:
2 монографии и 1 вузовский учебник (по математике)
и
23 статьи в математических журналах
(Успехи мат. наук, Труды института им. Стеклова, Мат. Заметки и т.д.)

А также – опубликовано 36 статей в биологических и информатических журналах (Молекулярная биология, Биофизика, Биохимия, FEMC, BMC, JBCB, inSB, ...)

Подготовлено 2 докторские и 3 кандидатские диссертации (все – физ.-мат. науки, «теоретические основы информатики», «биоинформатика»)

Ежегодно наши аспиранты и сотрудники делают доклады примерно на 4-х международных конференциях (математических, биологических, информатических)

За это время аспиранты и сотрудники приняли участие в выполнении: 25 грантов, 2 целевых грантов, 2 научных программ и 2 совместных тем по линии РАН-СНРС.

Лауреаты премии

«За лучшую публикацию в журнале Молекулярная биология» за 2005 год и премий некоторых зарубежных университетов

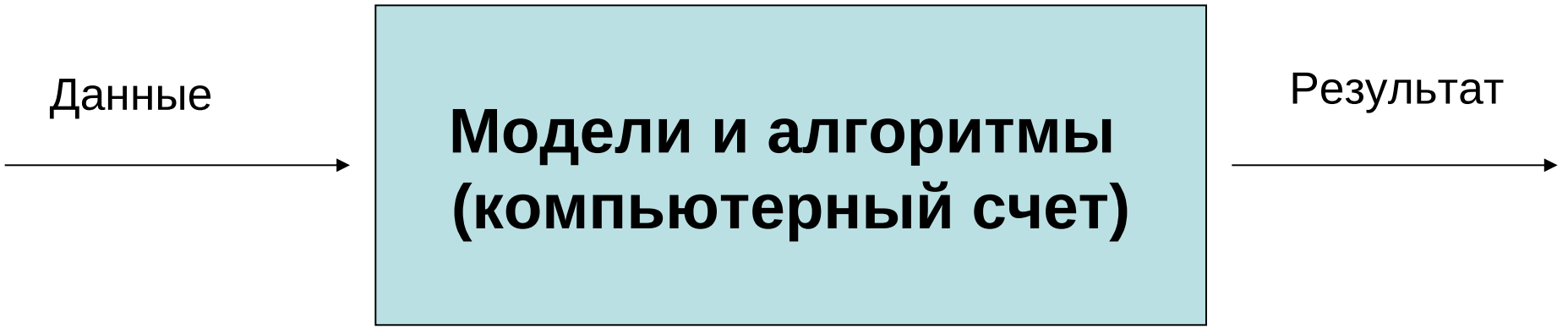
Тесно сотрудничаем с кафедрой «математической логики и теории алгоритмов» мех-мата МГУ, в частности, читаем там курс «Модели и алгоритмы в биоинформатике». Сотрудничаем с факультетом биоинформатики и биоинженерии МГУ, с факультетом ВМК.

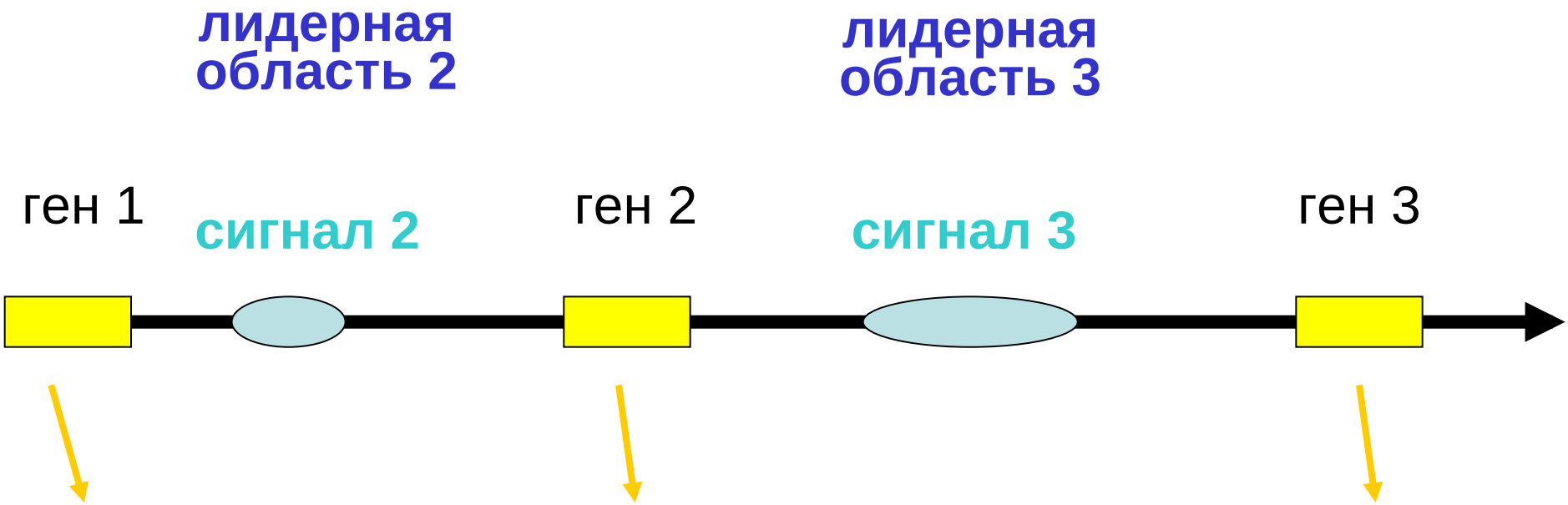
Регулярно ведем курсовые и дипломные работы, аспирантов. Сотрудничаем с аспирантурой Париж-7.
Включая оплачиваемую работу.

1) Проблемы **эффективности**;

2) **Модели** и **алгоритмы** основных молекулярных процессов в клетке: геномы бактерий, растений, водорослей и простейших .....

ДНК=геном – последовательность в 4х-буквенном алфавите {A,C,T,G}
с характерной длиной 3 миллиона – 6 миллиардов позиций. Каждая буква называется «**нуклеотид**».

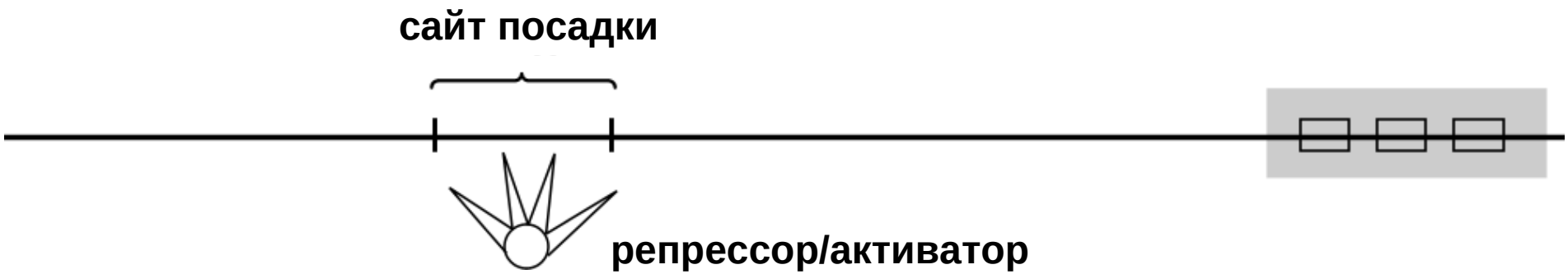

Данные
Модели и алгоритмы
(компьютерный счет)
Результат



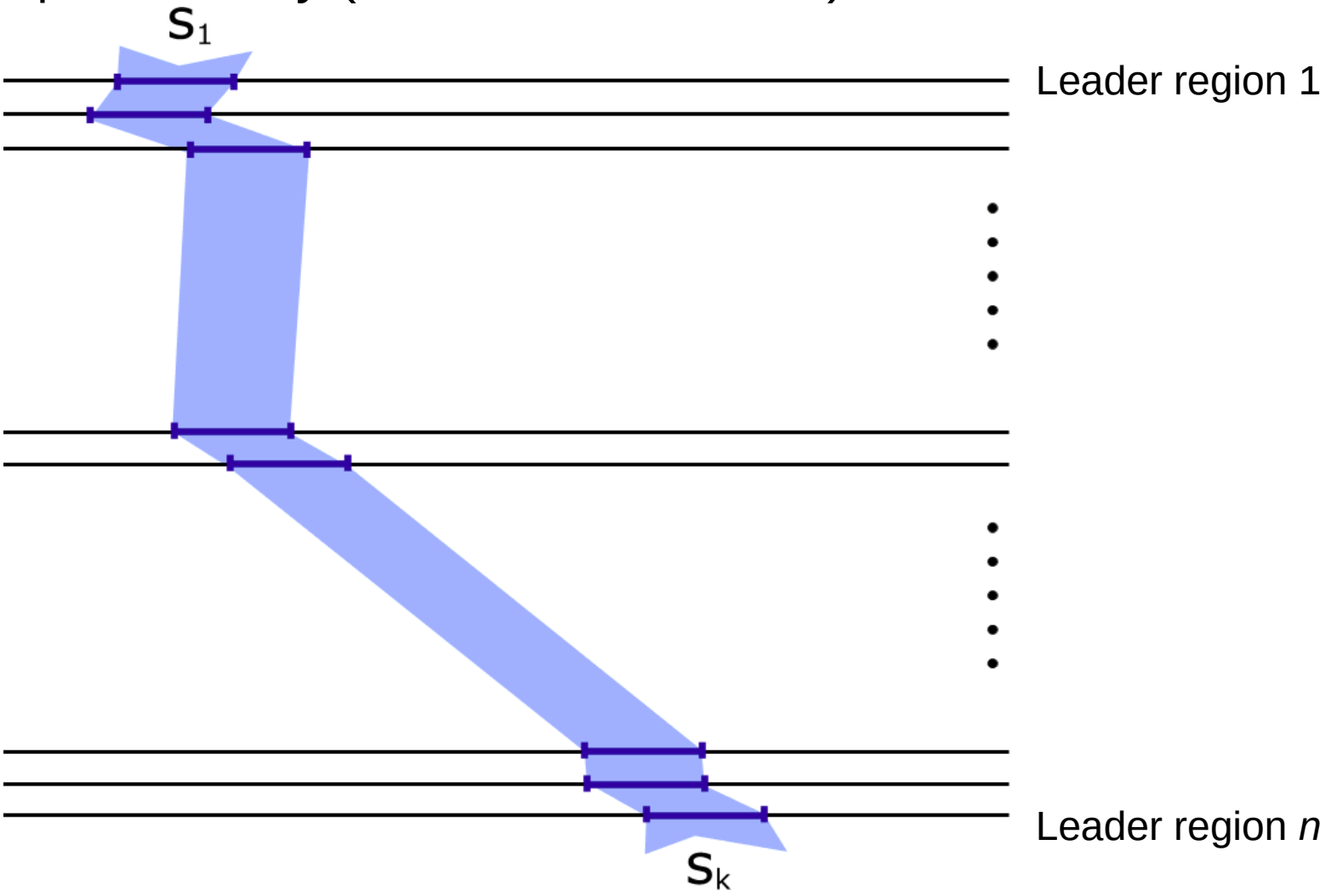
Ген считывается по сигналу из лидерной области!

Ген и сигнал эволюционируют!

Один из возможных типов сигналов (= регуляций):



Даны n последовательностей. **Задача**: найти **систему сайтов (=сигнал,мотив)** $s = \{s_1,...,s_k\}$, состоящую из сайтов $s_1,...,s_k$, где $k \leq n$. Все сайты имеют одинаковую длину. **Определяем качество системы** как сумму попарных близостей сайтов, составляющих систему (=качество сигнала).



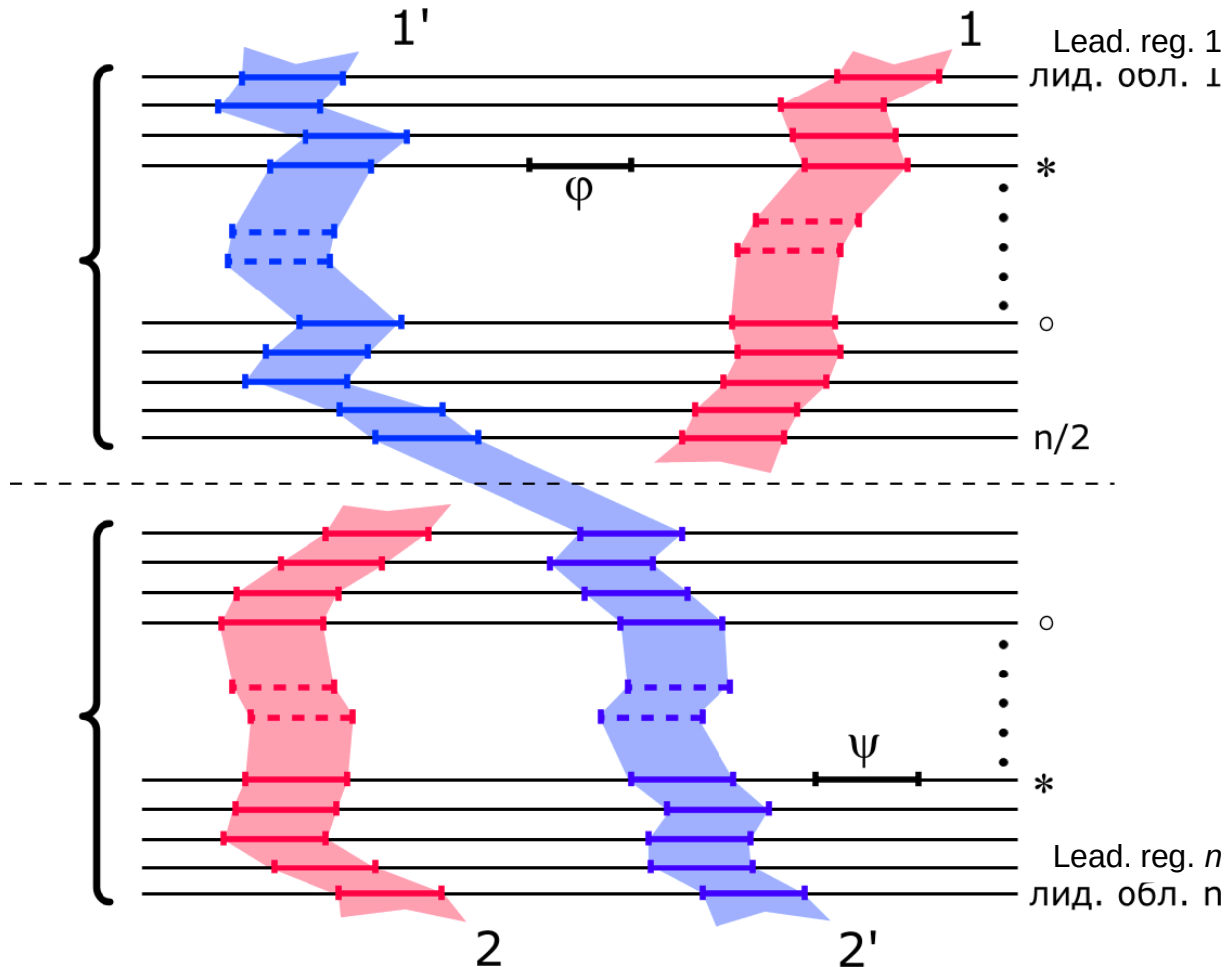
Ищем систему сайтов с максимальным значением качества, т.е. ищем минимум целевого функционала *F* в пространстве всех возможных систем:

$$F(s = \{s_1, \ldots, s_k\}) = \sum_{i,j=1}^{k} \rho(s_i, s_j)$$

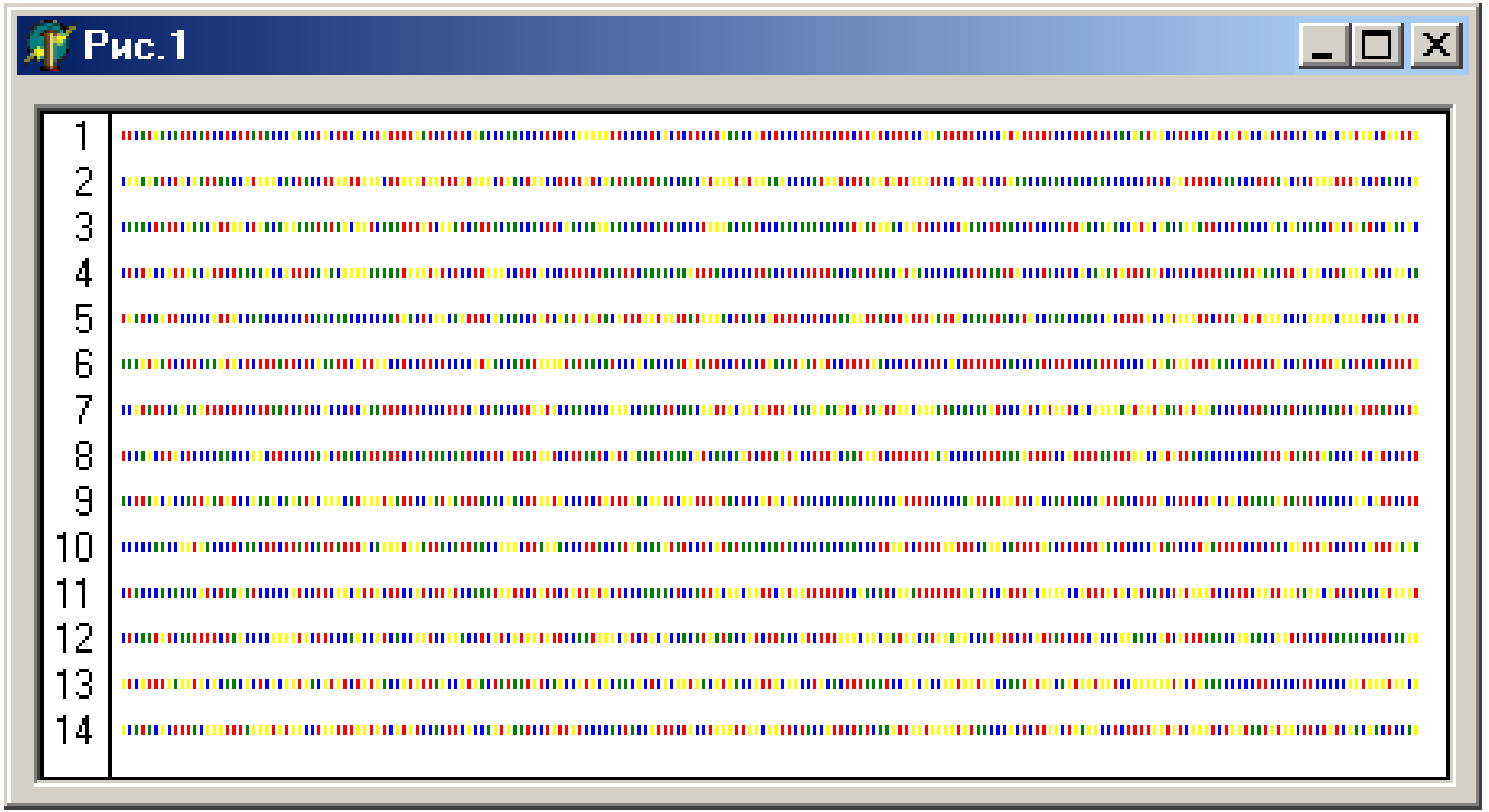
Идея нашего алгоритма. **Делим все последовательности на две примерно равные части** и **лучшую** систему в одной части объединяем с **лучшей** системой в другой части. Пусть $\Gamma_1(\varphi)$ – **лучшая система** в одной части **как функция от** φ **(и фиксирована последовательность *)**, а $\Gamma_2(\psi)$ – аналогичная система в другой части **как функция от** ψ .

Индуктивный шаг:

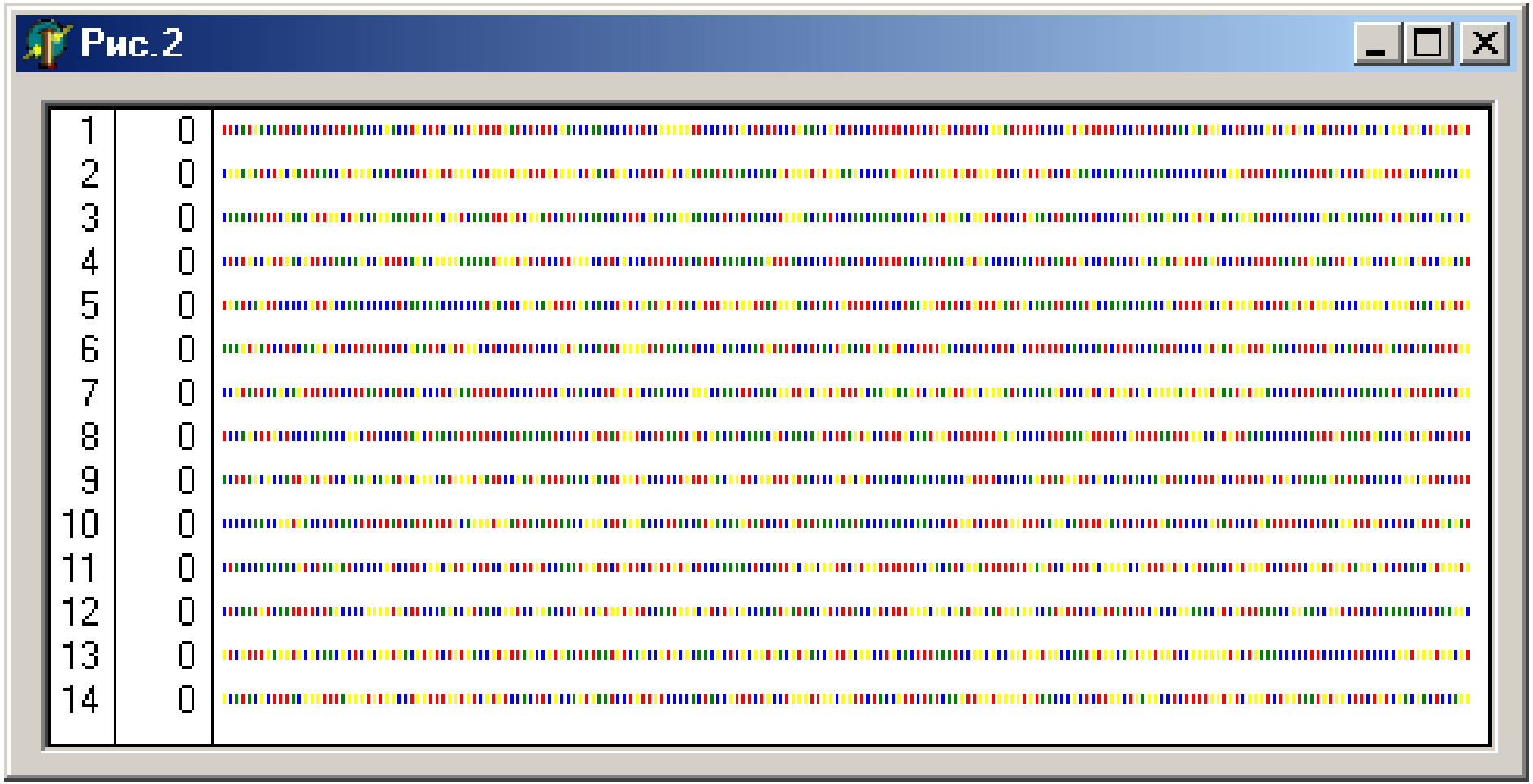
от $\Gamma_1(\bullet)$ и $\Gamma_2(\bullet)$ переходим к $\Gamma(\bullet)$ по правилу: Γ лучшая система $\Gamma_1(\varphi)+\Gamma_2(\psi)$, полученная **перебором всех** φ **и** ψ **в** *последовательностях



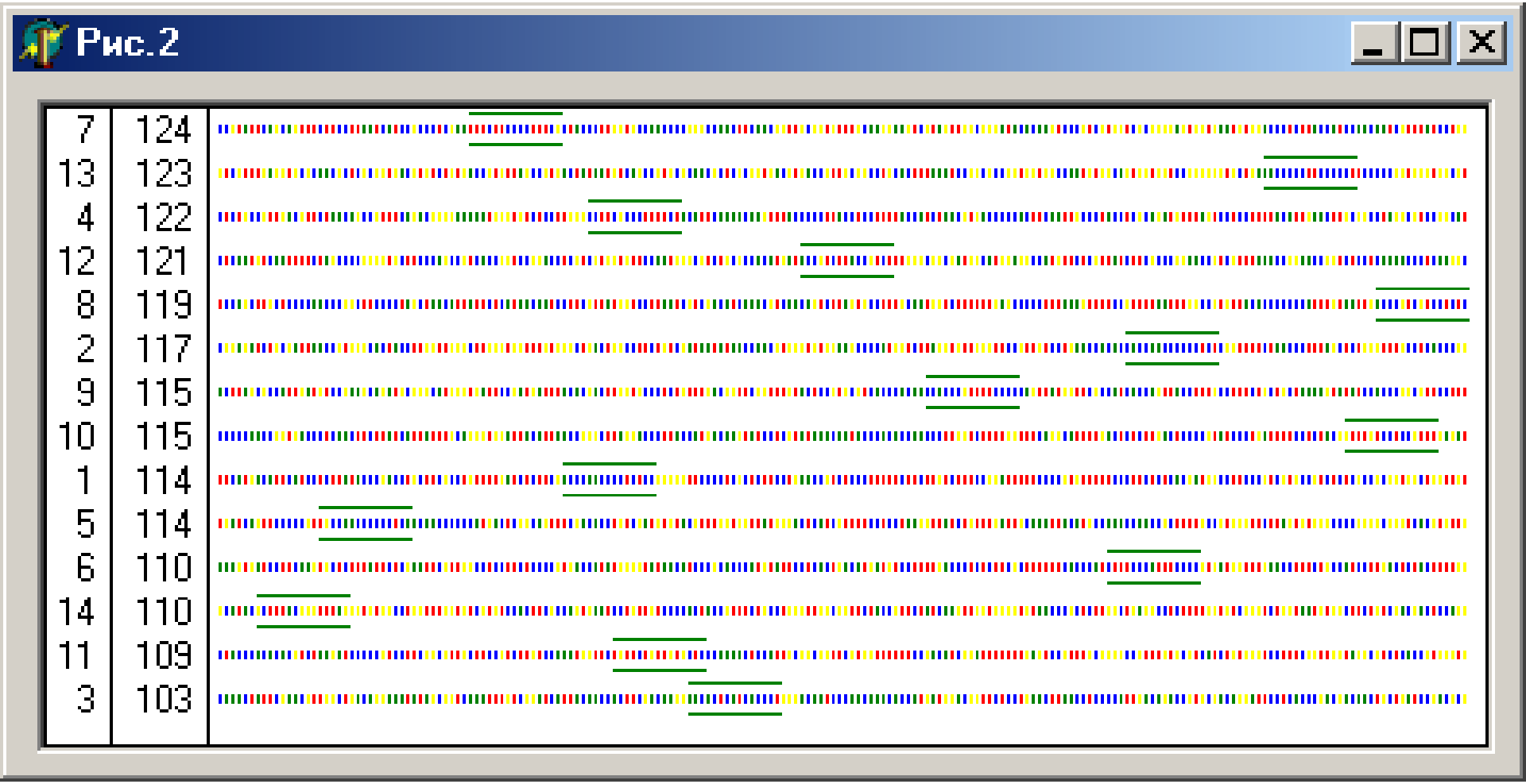
Пример. Даны n=14 последовательностей, каждая с длиной m=201; ищем систему сайтов с длиной 15



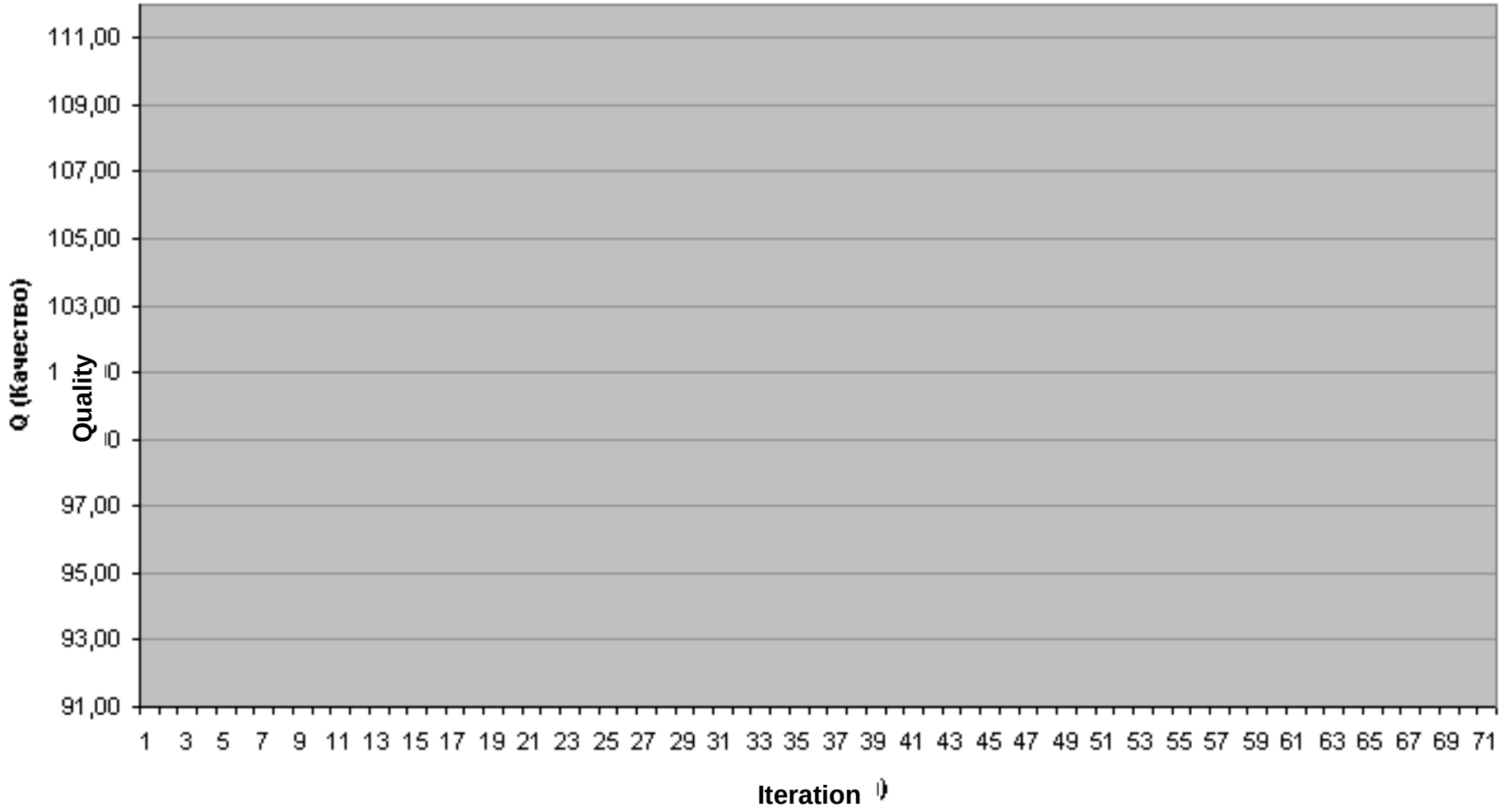
Работа алгоритма:



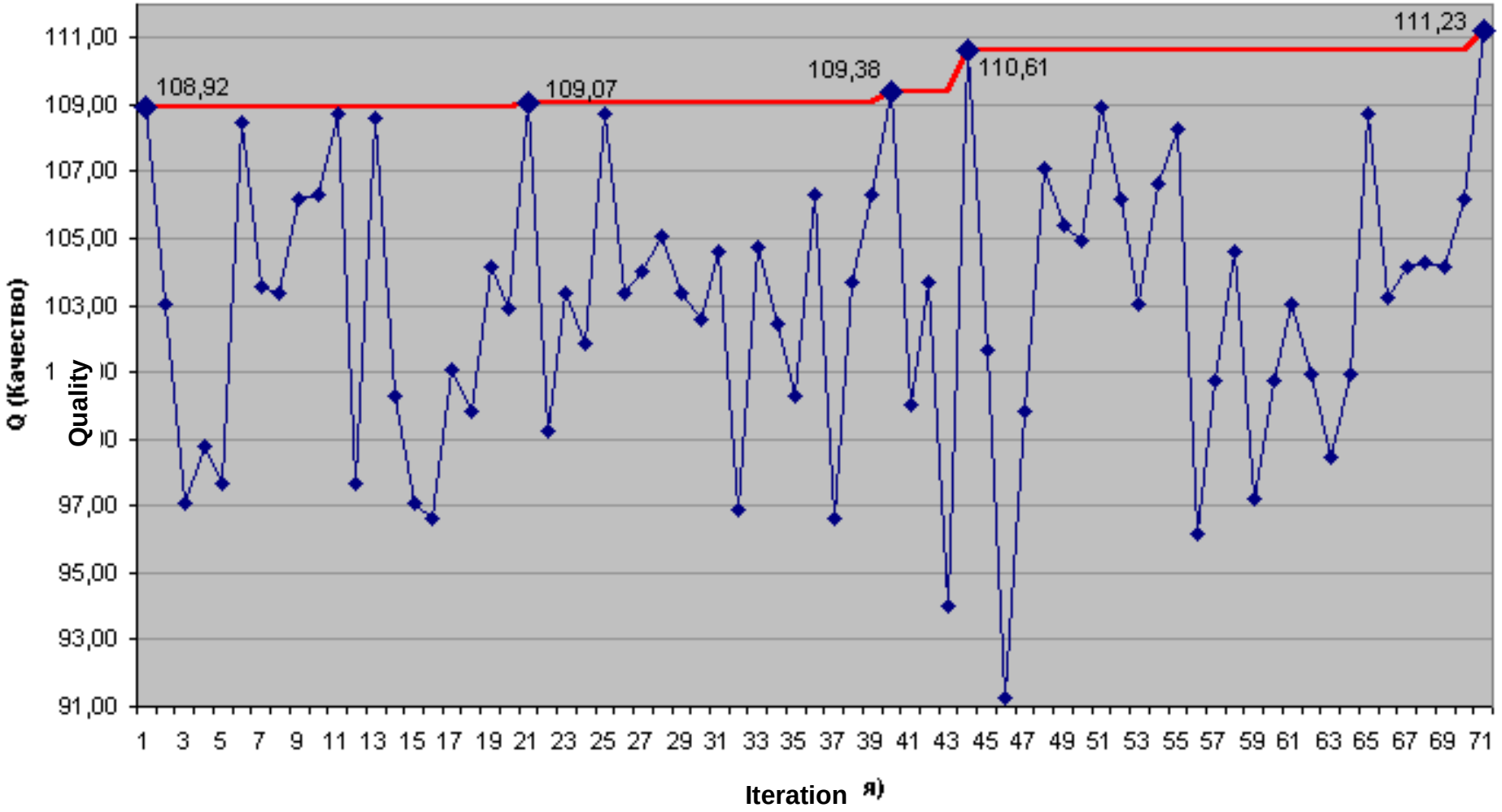
Результат работы алгоритма:



Качество потенциального сигнала растет в процессе счета:



Последовательное изменение качества сигнала в ходе алгоритма:



Параллельная реализация вычислительно трудоемких алгоритмов: поиск мультибоксового регуляторного сигнала в группе геномов

$P_0(0)\rightarrow$	$P_1(1)\rightarrow$	$P_2(2)\rightarrow$	$P_3(3)\rightarrow$	$P_4(4)\rightarrow$	$P_5(5)\rightarrow$	$P_6(6)\rightarrow$	$P_7(7)\rightarrow$	$P_8(59)\rightarrow$	$P_9(68)\rightarrow\ldots$
7734	9064	8872	**11468**	9314	8746	8696	9138	**9622**	8106
$\downarrow Q_{0,0}(11)$ 9972	$\downarrow Q_{1,0}(12)$ 10654	$\downarrow Q_{2,0}(14)$ 10776	$\downarrow Q_{3,0}(13)$ 9310	$\downarrow Q_{4,0}(8)$ 8348	$\downarrow Q_{5,0}(10)$ 12150	$\downarrow Q_{6,0}(9)$ 11756	$\downarrow Q_{7,0}(15)$ **10018**	$\downarrow Q_{8,0}(66)$ 9140	$\downarrow Q_{9,0}(75)$ **11472**
$\downarrow Q_{0,1}(18)$ 7860	$\downarrow Q_{1,1}(19)$ **11006**	$\downarrow Q_{2,1}(22)$ 9348	$\downarrow Q_{3,1}(21)$ 9584	$\downarrow Q_{4,1}(17)$ 9388	$\downarrow Q_{5,1}(16)$ 12056	$\downarrow Q_{6,1}(20)$ **12628**	$\downarrow Q_{7,1}(23)$ 8408	$\downarrow Q_{8,1}(76)$ 8996	…
$\downarrow Q_{0,2}(26)$ 8372	$\downarrow Q_{1,2}(27)$ 10522	$\downarrow Q_{2,2}(31)$ 8612	$\downarrow Q_{3,2}(28)$ 8922	$\downarrow Q_{4,2}(25)$ 8602	$\downarrow Q_{5,2}(24)$ 12040	$\downarrow Q_{6,2}(30)$ 11200	$\downarrow Q_{7,2}(29)$ 8996	…	
$\downarrow Q_{0,3}(34)$ 9784	$\downarrow Q_{1,3}(38)$ 8394	$\downarrow Q_{2,3}(39)$ 11082	$\downarrow Q_{3,3}(35)$ 8734	$\downarrow Q_{4,3}(33)$ 9022	$\downarrow Q_{5,3}(32)$ 11942	$\downarrow Q_{6,3}(36)$ 11626	$\downarrow Q_{7,3}(37)$ 9416		
$\downarrow Q_{0,4}(42)$ 10188	$\downarrow Q_{1,4}(44)$ 8716	$\downarrow Q_{2,4}(47)$ 9928	$\downarrow Q_{3,4}(43)$ 9238	$\downarrow Q_{4,4}(41)$ 9284	$\downarrow Q_{5,4}(40)$ 11448	$\downarrow Q_{6,4}(45)$ 9262	$\downarrow Q_{7,4}(46)$ 8744		
$\downarrow Q_{0,5}(50)$ 10100	$\downarrow Q_{1,5}(52)$ 9546	$\downarrow Q_{2,5}(54)$ 9564	$\downarrow Q_{3,5}(51)$ 8958	$\downarrow Q_{4,5}(49)$ **9490**	$\downarrow Q_{5,5}(48)$ 12164	$\downarrow Q_{6,5}(53)$ 11526	$\downarrow Q_{7,5}(55)$ 8448		
$\downarrow Q_{0,6}(58)$ **12850**	$\downarrow Q_{1,6}(60)$ 10982	$\downarrow Q_{2,6}(63)$ 11702	×	$\downarrow Q_{4,6}(57)$ 8634	$\downarrow Q_{5,6}(56)$ **12668**	$\downarrow Q_{6,6}(61)$ 12104	$\downarrow Q_{7,6}(62)$ 8358		
$\downarrow Q_{0,7}(67)$ 12054	×	$\downarrow Q_{2,7}(70)$ **11766**		$\downarrow Q_{4,7}(65)$ 9088	$\downarrow Q_{5,7}(64)$ 12286	$\downarrow Q_{6,7}(71)$ 12058	$\downarrow Q_{7,7}(69)$ 8714		
$\downarrow Q_{0,8}(74)$ 8918		$\downarrow Q_{2,8}(78)$ 10306		$\downarrow Q_{4,8}(73)$ 8112	$\downarrow Q_{5,8}(72)$ 11632	×	$\downarrow Q_{7,8}(77)$ 8492		
…		…		…	…		…		

Пример для n=45, m=201, 8 CPU

«Однобоксовый» сигнал:
- полный перебор $O(m^n)$
- наш алгоритм $O(n^2m^3)$

«Двухбоксовый» сигнал:
- полный перебор $O(m^n d^n)$
- наш алгоритм $O(n^2m^3d^3)$

(n – число последовательностей, m – максимальная длина, d – интервал расстояний между боксами сигнала)

Волновая вычислительная схема на двумерной ε-сети перестановок мощностью порядка n^2 (в полном пространстве $n!$ перестановок):

1) отсутствует жёсткая привязка к числу процессоров кластера
2) линейный рост производительности от числа доступных процессоров в широком диапазоне (проверено на МВС-1000М МСЦ, до 512 CPU)

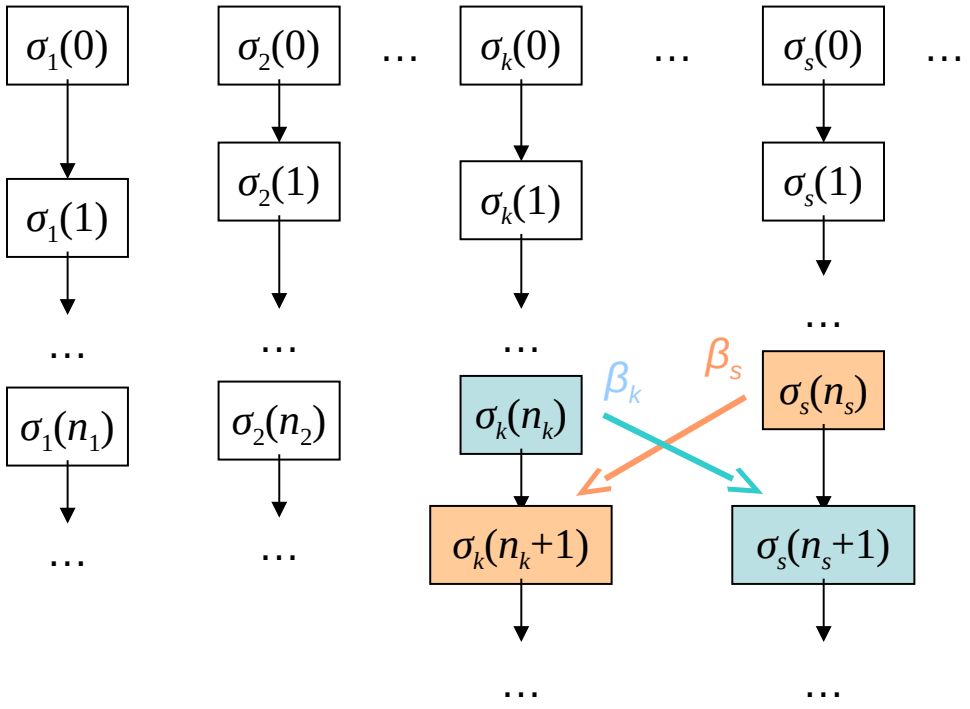
Wavelike computation scheme

Using 2D queue of permutations (*P*,*Q*) instead of straight one

$P_0(0)\rightarrow$	$P_1(1)\rightarrow$	$P_2(2)\rightarrow$	$P_3(3)\rightarrow$	$P_4(4)\rightarrow$	$P_5(5)\rightarrow$	$P_6(6)\rightarrow$	$P_7(7)\rightarrow$	$P_8(59)\rightarrow$	$P_9(68)\rightarrow$	$P_{10}(79)\rightarrow$...
175.8	206.0	201.6	260.6	211.7	198.8	197.6	207.7	218.7	184.2	214.2
$\downarrow Q_{0,0}(11)$	$\downarrow Q_{1,0}(12)$	$\downarrow Q_{2,0}(14)$	$\downarrow Q_{3,0}(13)$	$\downarrow Q_{4,0}(8)$	$\downarrow Q_{5,0}(10)$	$\downarrow Q_{6,0}(9)$	$\downarrow Q_{7,0}(15)$	$\downarrow Q_{8,0}(66)$	$\downarrow Q_{9,0}(75)$	...
226.6	242.1	244.9	211.6	189.7	276.1	267.2	227.7	207.7	260.7	
$\downarrow Q_{0,1}(18)$	$\downarrow Q_{1,1}(19)$	$\downarrow Q_{2,1}(22)$	$\downarrow Q_{3,1}(21)$	$\downarrow Q_{4,1}(17)$	$\downarrow Q_{5,1}(16)$	$\downarrow Q_{6,1}(20)$	$\downarrow Q_{7,1}(23)$	$\downarrow Q_{8,1}(76)$	...	
178.6	250.1	212.5	217.8	213.3	274.0	287.0	191.0	204.5		
$\downarrow Q_{0,2}(26)$	$\downarrow Q_{1,2}(27)$	$\downarrow Q_{2,2}(31)$	$\downarrow Q_{3,2}(28)$	$\downarrow Q_{4,2}(25)$	$\downarrow Q_{5,2}(24)$	$\downarrow Q_{6,2}(30)$	$\downarrow Q_{7,2}(29)$	...		
190.3	239.1	195.7	202.8	195.5	273.6	254.5	204.5			
$\downarrow Q_{0,3}(34)$	$\downarrow Q_{1,3}(38)$	$\downarrow Q_{2,3}(39)$	$\downarrow Q_{3,3}(35)$	$\downarrow Q_{4,3}(33)$	$\downarrow Q_{5,3}(32)$	$\downarrow Q_{6,3}(36)$	$\downarrow Q_{7,3}(37)$			
222.4	190.8	251.9	198.5	205.0	271.4	264.2	214.0			
$\downarrow Q_{0,4}(42)$	$\downarrow Q_{1,4}(44)$	$\downarrow Q_{2,4}(47)$	$\downarrow Q_{3,4}(43)$	$\downarrow Q_{4,4}(41)$	$\downarrow Q_{5,4}(40)$	$\downarrow Q_{6,4}(45)$	$\downarrow Q_{7,4}(46)$			
231.5	198.1	225.6	210.0	211.0	260.2	210.5	198.8			
$\downarrow Q_{0,5}(50)$	$\downarrow Q_{1,5}(52)$	$\downarrow Q_{2,5}(54)$	$\downarrow Q_{3,5}(51)$	$\downarrow Q_{4,5}(49)$	$\downarrow Q_{5,5}(48)$	$\downarrow Q_{6,5}(53)$	$\downarrow Q_{7,5}(55)$			
229.5	217.0	217.4	203.6	215.7	276.5	262.0	192.0			
$\downarrow Q_{0,6}(58)$	$\downarrow Q_{1,6}(60)$	$\downarrow Q_{2,6}(63)$	=====	$\downarrow Q_{4,6}(57)$	$\downarrow Q_{5,6}(56)$	$\downarrow Q_{6,6}(61)$	$\downarrow Q_{7,6}(62)$			
292.0	249.6	266.0		196.2	287.9	275.1	190.0			
$\downarrow Q_{0,7}(67)$	=====	$\downarrow Q_{2,7}(70)$		$\downarrow Q_{4,7}(65)$	$\downarrow Q_{5,7}(64)$	$\downarrow Q_{6,7}(71)$	$\downarrow Q_{7,7}(69)$			
274.0		267.6		206.5	279.2	274.0	198.0			
$\downarrow Q_{0,8}(74)$		$\downarrow Q_{2,8}(78)$		$\downarrow Q_{4,8}(73)$	$\downarrow Q_{5,8}(72)$	=====	$\downarrow Q_{7,8}(77)$			
202.7		234.2		184.4	264.4		193.0			
...		...		...	...		...			

n=45,
m=201,
l=15,
8 CPU's

Параллельная реализация вычислительно трудоемких алгоритмов: реконструкция эволюции регуляторного сигнала в группе геномов



Индивидуальные режимы охлаждения

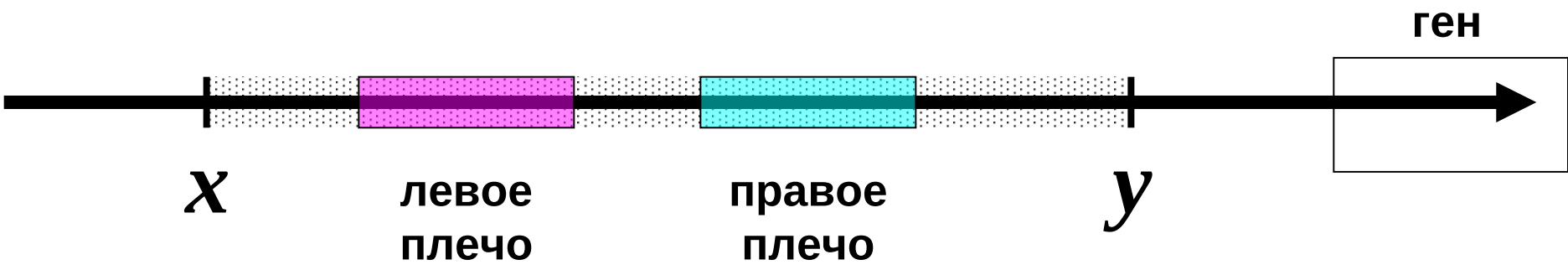
$$\beta_i(n) = \frac{C}{1 + B(i-1)} \log^p(n+1)$$

Периодический обмен параметрами охлаждения между находящимися в окрестности различных локальных или условных минимумов цепями с разной температурой способствует выходу из оврагов и локальных минимумов поверхности отклика.

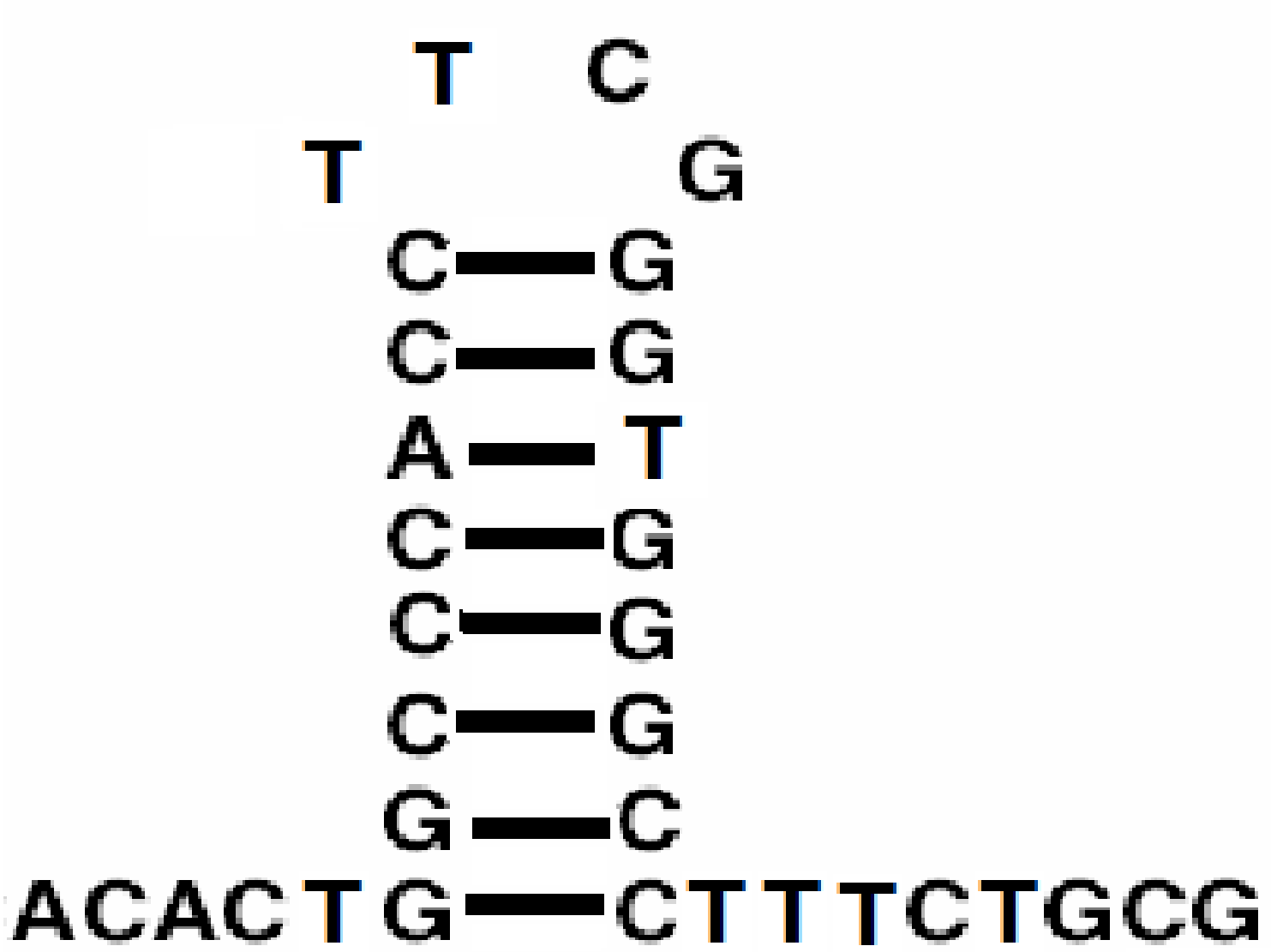
Усовершенствованная **параллельная схема аннилинга** MC^3 (= Metropolis-Coupled Markov Chain Monte-Carlo):

1) лучшее покрытие множества минимальных конфигураций
2) меньшая зависимость от выбранной начальной точки
3) более быстрая сходимость к одному из предполагаемых абсолютных минимумов функционала «энергии»

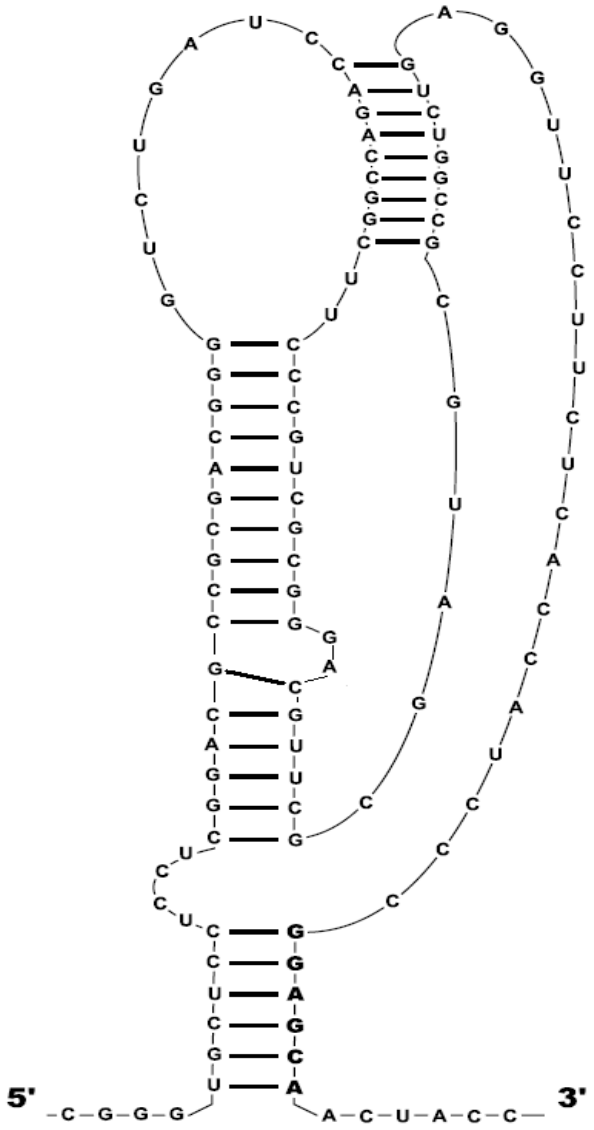
Показана лидерная область перед геном,
в ней «**окно**» с концами *x* и *y*,
а в окне образуются «**спирали**»

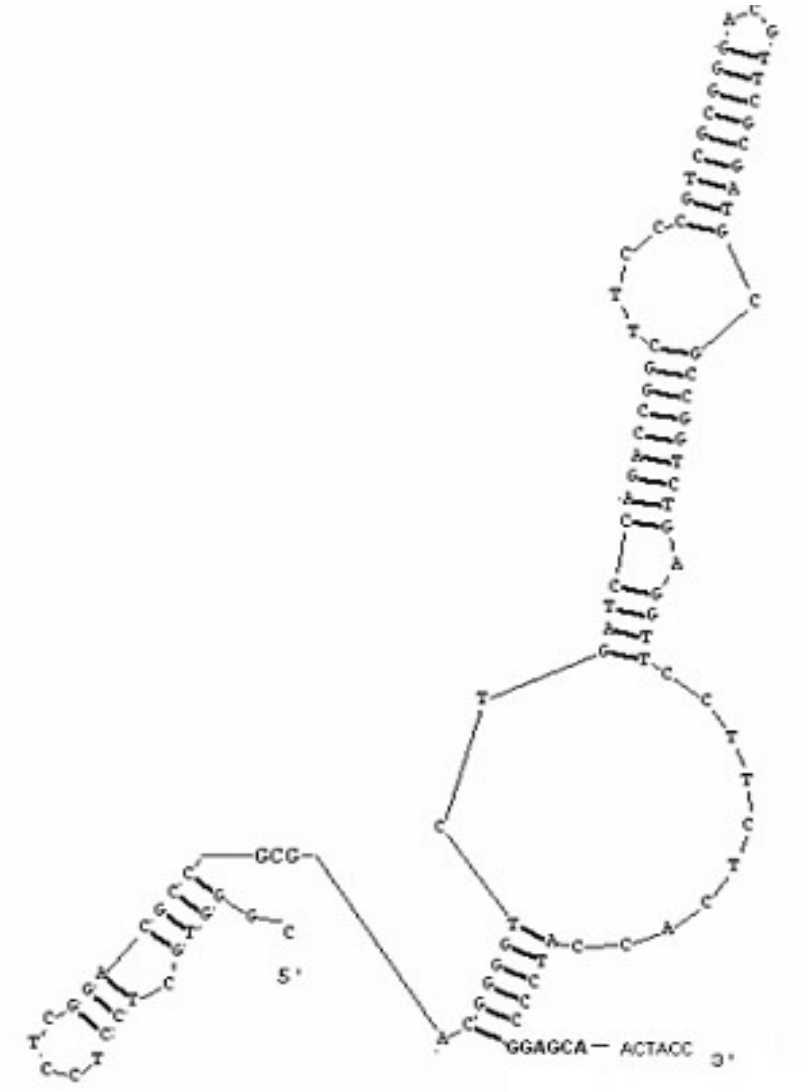


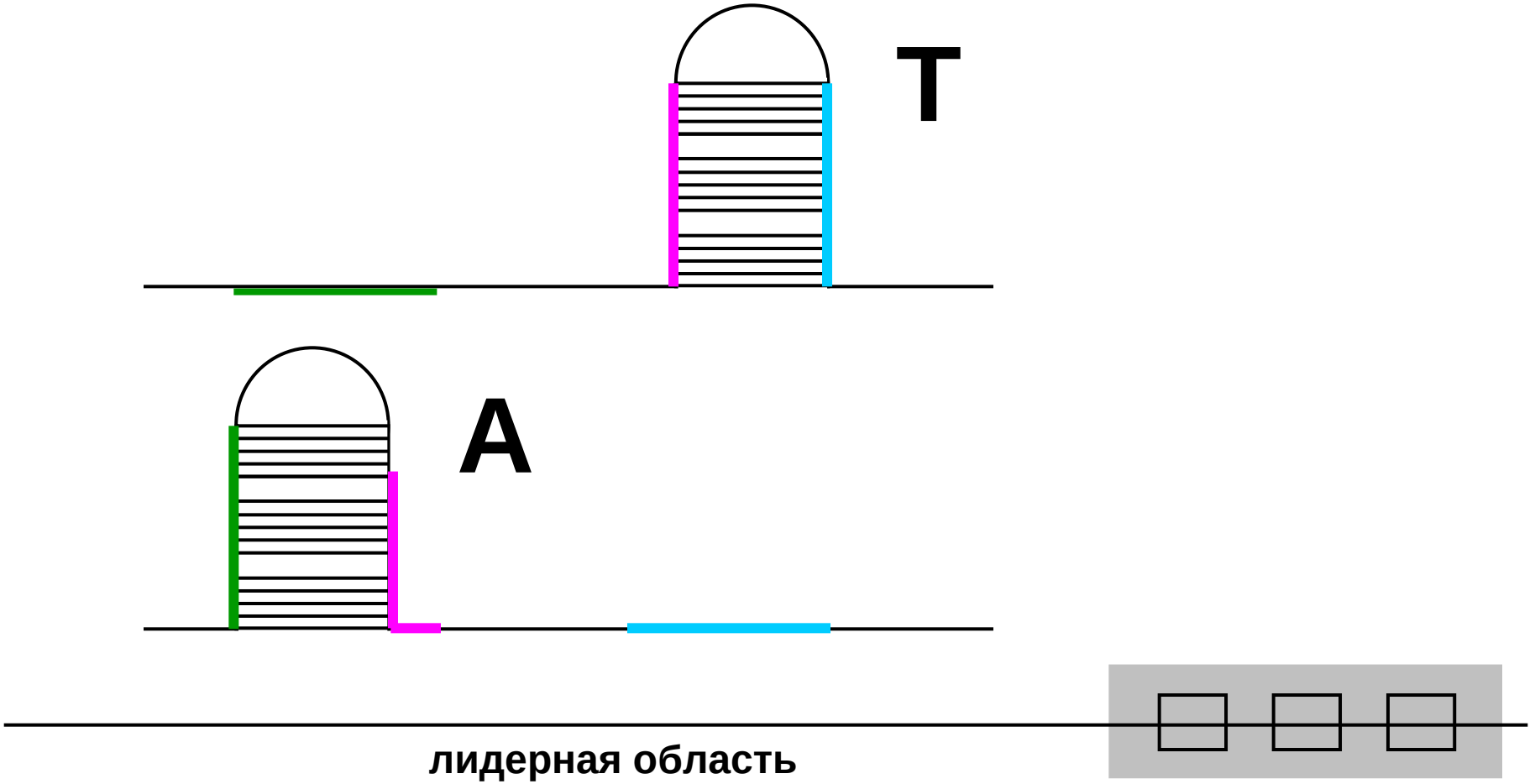
«Спираль» с «плечами»,
склеиваются G с **C** и **A** с T:



Реальные еще очень простые вторичные структуры (=наборы спиралей):

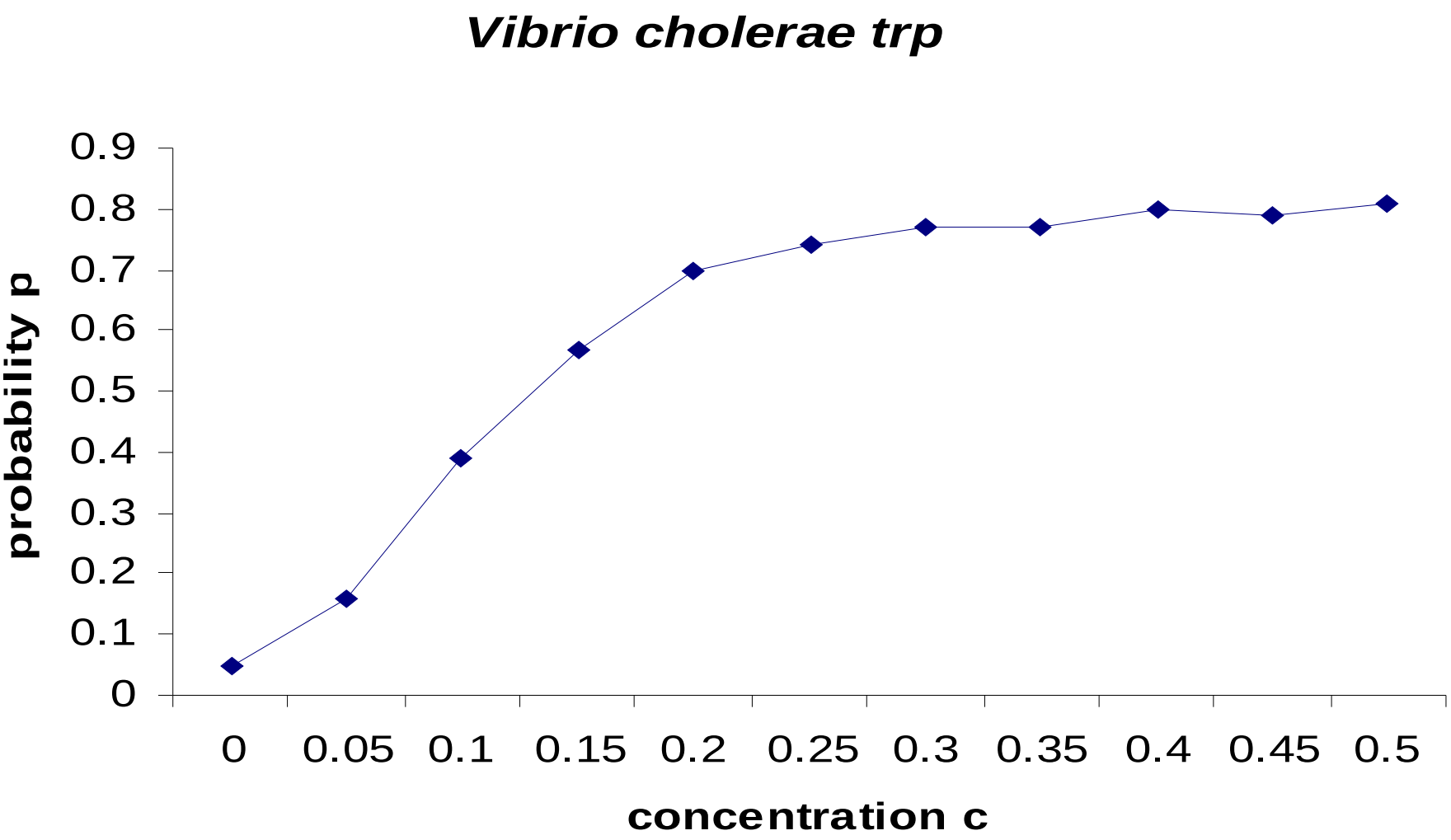






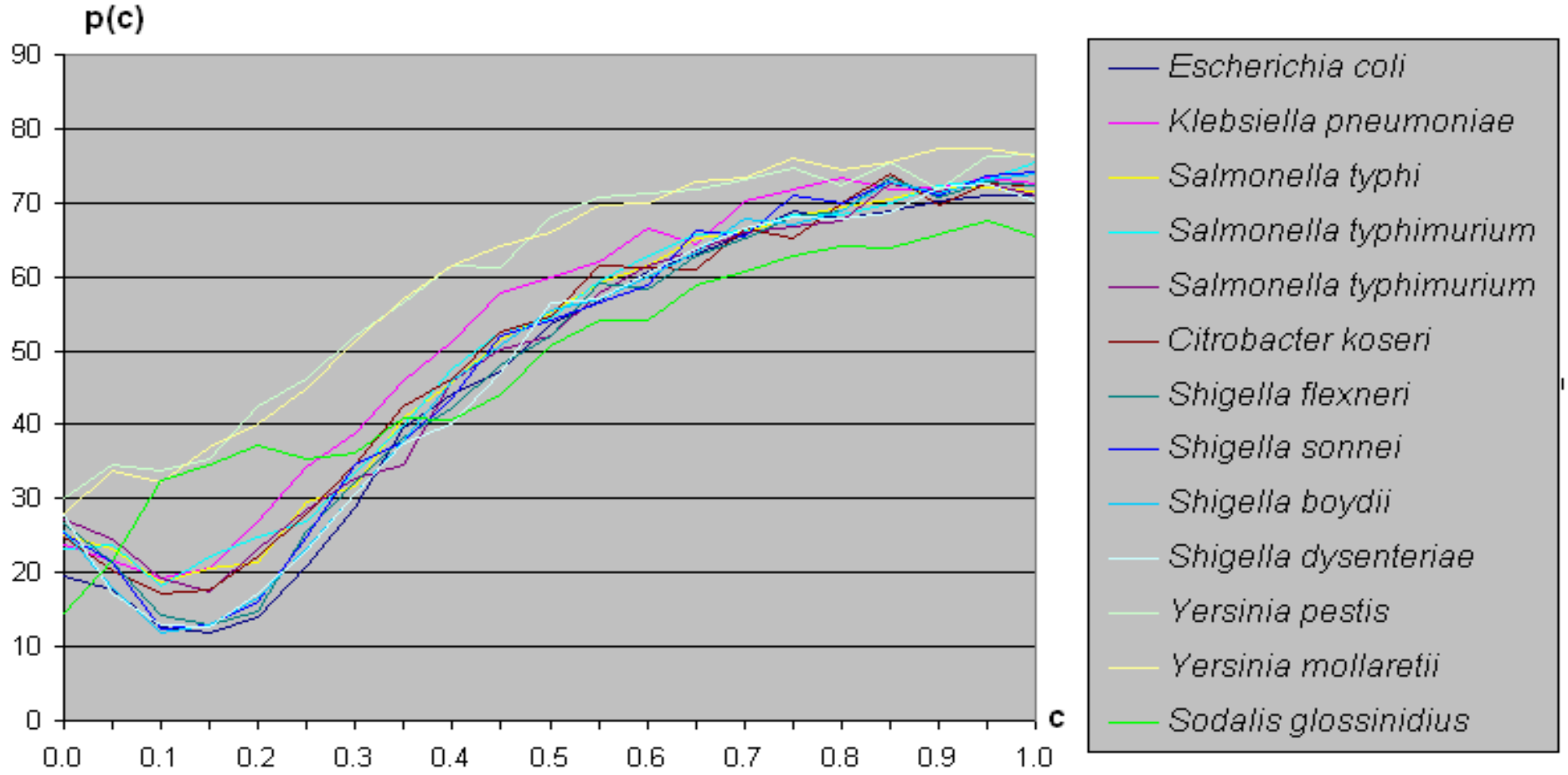
Два состояния сигнала. Результат определяется тем, какая из двух **альтернативных** вторичных структур образуется: «**T**» или «**A**»

Результат одной нашей моделей регуляции:



Примеры результатов счета в этой модели

Мы считали функцию $p=p(c)$ для практически всех лидерных областей аминокислотных оперонов и аминоацил-тРНК синтетаз. Имеется **высокое согласие с экспериментом**, с одной стороны, и **предсказание многих новых случаев такой регуляции**, с другой стороны. Здесь показаны *thrA* опероны у гамма-протеобактерий.

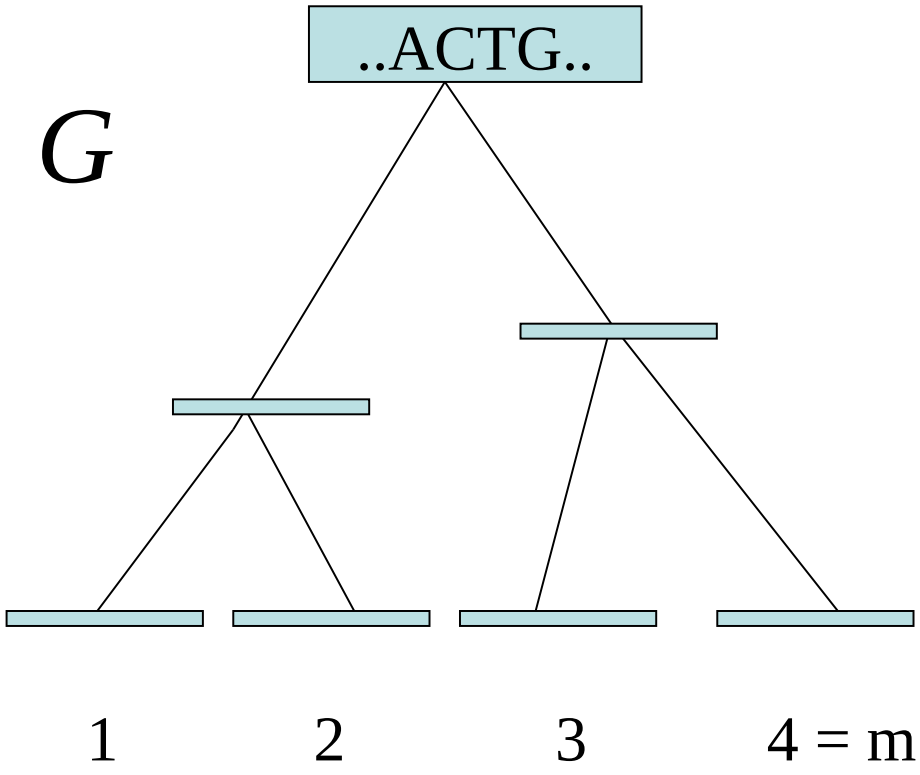


Два основных направления
нашей работы в Биоинформатике:

1) **Модели и алгоритмы регуляции генов**,

2) **Модели и алгоритмы эволюции этих регуляций (=сигналов)**

Дано дерево G , у которого длины ребер соответствуют времени переходу от предка к потомку.
Даны современные последовательности

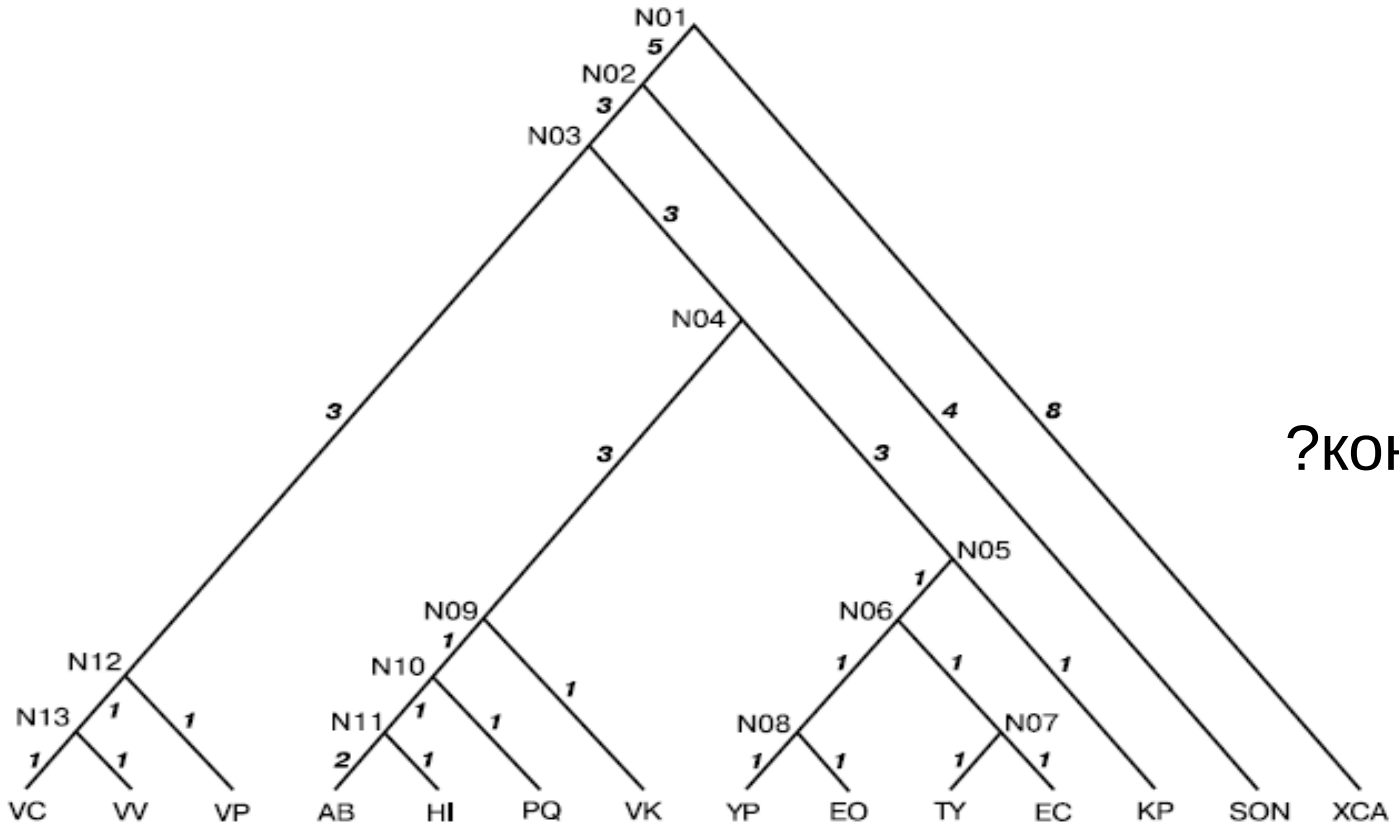


Ищем все **предковые последовательности**

Иногда **ищется** и само дерево G : тогда **даны** только современные последовательности.

Эти заданные последовательности – организмы, виды, гены, белки, сигналы

Классическая аттенюаторная регуляция биосинтеза треонина у гамма-протеобактерий



?конфигурация σ

VC = *Vibrio cholerae*, VV = *Vibrio vulnificus*, VP = *Vibrio parahaemolyticus*,
AB = *Actinobacillus actinomycetemcomitans*, HI = *Haemophylus influenzae*,
PQ = *Mannheimia haemolytica*, VK = *Pasterella multocida*, YP = *Yersinia pestis*,
EO = *Erwinia carotovora*, TY = *Salmonella typhi* , XCA = *Xanthomonas campestris*,
EC = *Escherichia coli*, KP = *Klebsiella pneumoniae*, SON = *Shewanella oneidensis*

Наша модель эволюции сигнала:

$$H(\sigma) = H_1(\sigma) + H_2(\sigma) \to gl\min$$

Такая функция минимизируется с помощью алгоритма аннилинга. На каждом его шаге текущая конфигурация σ заменяется на новую $\tilde{\sigma}$ из определенного списка возможностей с вероятностью

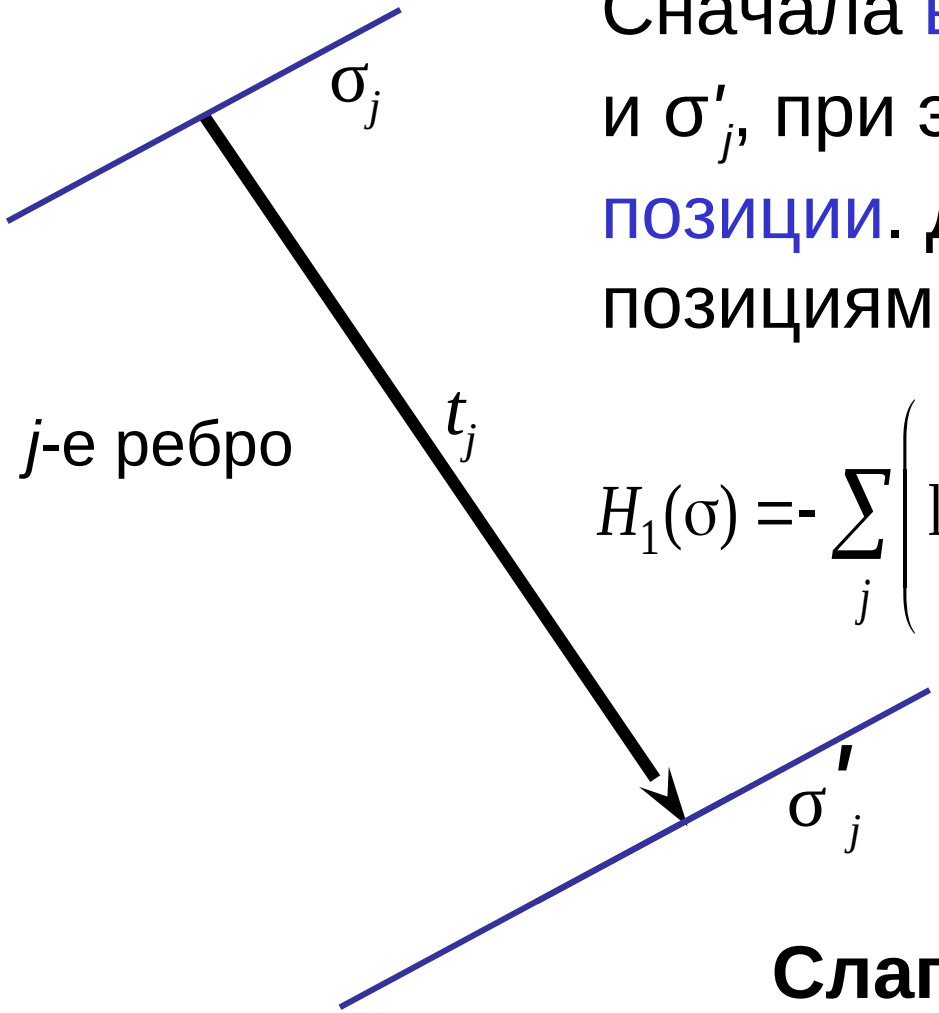
$$q(\sigma,\tilde{\sigma}) = \exp\left\{-\beta_m \cdot \left[H(\sigma) - H(\sigma)\right]^+\right\}$$

или остается прежней с вероятностью $(1-q)$.

Нами доказана сходимость к **глобальному min** при условии

$$\lim_{m\to\infty} \frac{\log m}{\beta_m} > const$$

Показано одно ребро от некоторой конфигурации σ. На этом ребре за время t_j происходят: замены букв со скоростями R, вставки букв и делеции букв.

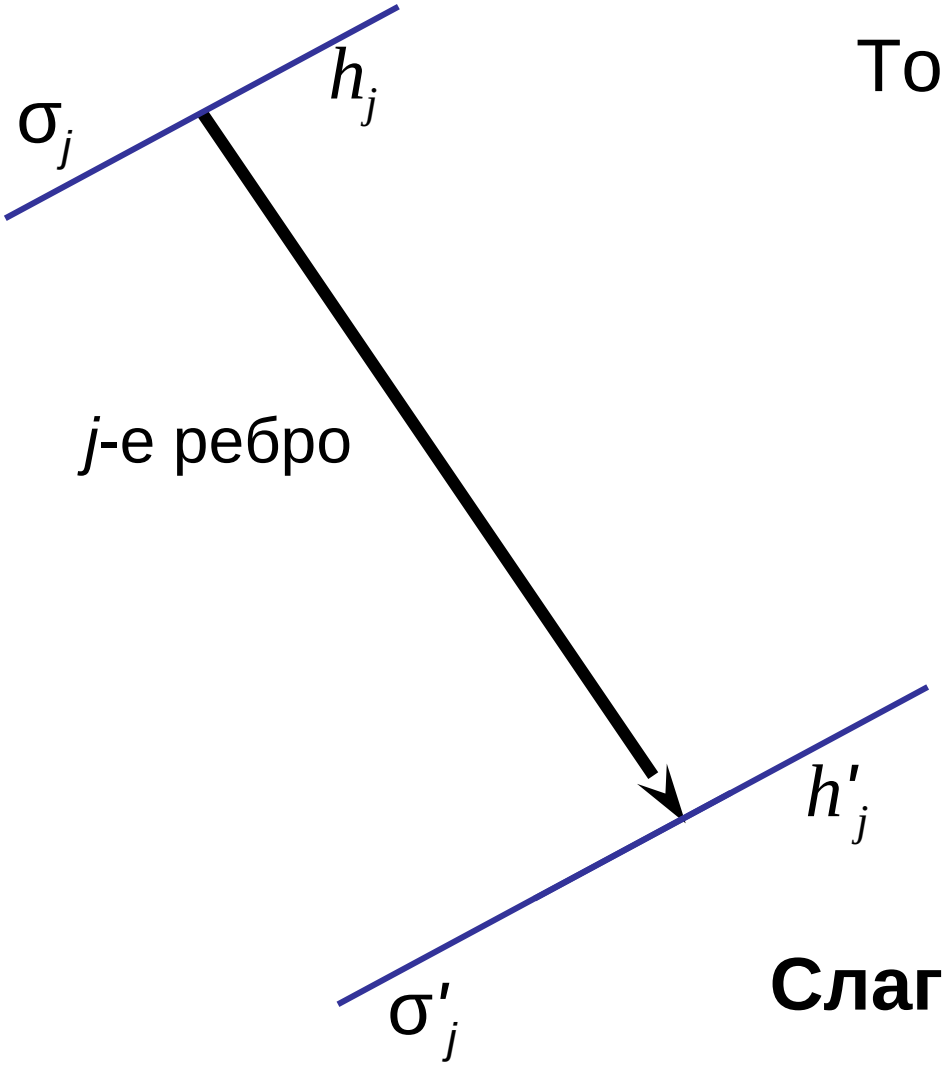


Сначала выравниваем позиции у σ_j и σ'_j, при этом возникают пустые позиции. Длины участков с пустыми позициями обозначим l_{jm}. Тогда:

$$H_1(\sigma) = -\sum_j \left(\ln \prod_{i=1}^{n_j}{}' \left(e^{\gamma_i t_j R} \right) \left(\bar{\sigma}_{ji}, \bar{\sigma}'_{ji} \right) - 10 \cdot \sum_m \left(l_{j,m} + 1 \right) \right)$$

Слагаемое $H_1(\sigma)$ в функции H

Показано одно ребро от конфигурации σ. На этом ребре произошел переход от вторичной структуры h_j в σ_j к вторичной структуре h'_j в σ'_j.



Тогда:

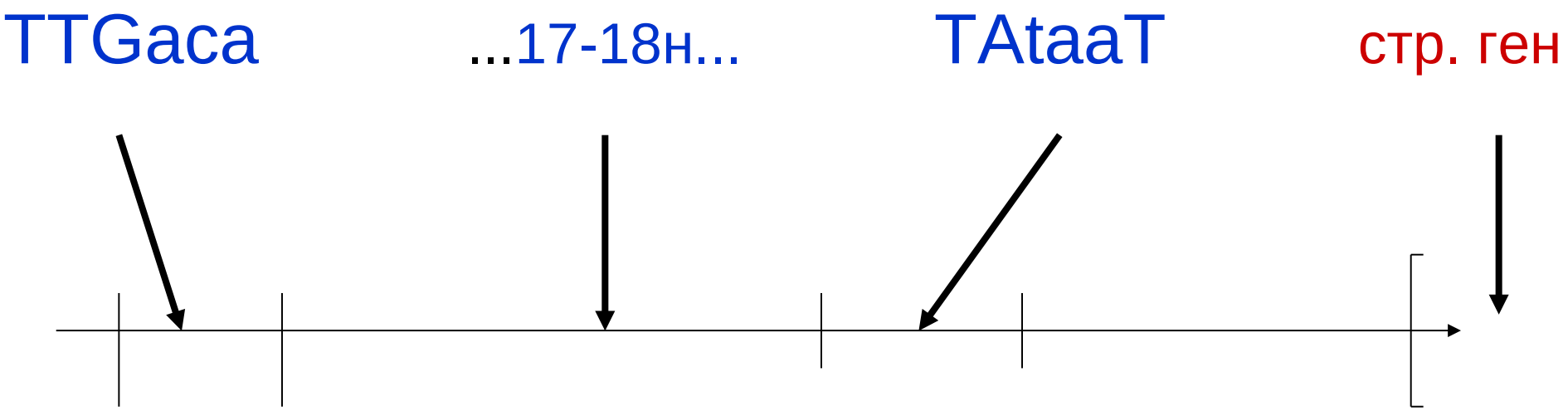
$$H_2(\sigma) = -\sum_j \Phi\left(h_j, h'_j\right)$$

Слагаемое $H_2(\sigma)$ в функции H

Решение (фрагмент): эволюция предкового сигнала

```
gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc      N01: H3=-29.2
gGTTGGGGCGGGCTgctgtactcaaaaaattttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      N02: H3=-51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      N03: H3=-45.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa      N12: H3=-61.3
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAATAaGATACAGGCCTTTTTTTa      N13: H3=-47.5
tGTTGGGGCAGGCTgctgagcgcaaaatttcacAAAAAAGGCCTGTATCCCAACcGATACAGGCCTTTTTTta      VC:  Σ=-234.3

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc      N01: H3=-29.2
gGTTGGGGCGGGCTgctgtactcaaaaaattttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      N02: H3=-51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      N03: H3=-45.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa      N12: H3=-61.3
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAATAaGATACAGGCCTTTTTTTa      N13: H3=-61.3
TGTTGGGGCAGGCTgctgagcgaaagaacaaatttcAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa   VV:  Σ=-248.1

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc      N01: H3=-29.2
gGTTGGGGCGGGCTgctgtactcaaaaaattttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      N02: H3=-51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      N03: H3=-45.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa      N12: H3=-57.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcacAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTta      VP:  Σ=-182.6

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc      N01: H3=-29.2
gGTTGGGGCGGGCTgctgtactcaaaaaattttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      N02: H3=-51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      N03: H3=-39.1
TGatGGTGCGGGCTgatgcgcacaagaaaaatcAGAAAAAAGCCCGCACCCAacaaaaTGCGGGCTTTTTTTTa      N04: H3=-24.6
aGAtgGTGCGGGTTagtgctgacaaaaaaaatgaacAAAAAACCCGCACTCaacaaaaAGCGGGTTTTTTtata      N09: H3=-39.0
aaTGGTGCGGGTTagtactggcaaaaaaaatgaacAAAAAACCCGCAaCTCAactaaaAGCGGGTTTTTTtata      N10: H3=-51.0
aaTGGTGCGGGTTagtacggcaaaaaaaagaaacAAAAAACCCGCAaCTCAactgaaAGCGGGTTTTTTtata      N11: H3=-6.2
aaTGGGGCGGGctagtgcgttgaagaatagaattcatGAACCCGCaTTTCCCGAGaGCGGGTTTttttatg         AB:  Σ=-240.5

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc      N01: H3=-29.2
gGTTGGGGCGGGCTgctgtactcaaaaaattttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      N02: H3=-51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      N03: H3=-39.1
TGatGGTGCGGGCTgatgcgcacaagaaaaatcAGAAAAAAGCCCGCACCCAacaaaaTGCGGGCTTTTTTTTa      N04: H3=-24.6
aGAtgGTGCGGGTTagtgctgacaaaaaaaatgaacAAAAAACCCGCACTCaacaaaaAGCGGGTTTTTTtata      N09: H3=-39.0
aaTGGTGCGGGTTagtactggcaaaaaaaatgaacAAAAAACCCGCAaCTCAactaaaAGCGGGTTTTTTtata      N10: H3=-51.0
aaTGGTGCGGGTTagtacggcaaaaaaaagaaacAAAAAACCCGCAaCTCAactgaaAGCGGGTTTTTTtata      N11: H3=-35.0
aaTGGTGCGGGTTagtgcagcaaaaacaagatacAGAAAACCCGCGATTCAactGAATaGCGGGTTTTTTtata      HI:  Σ=-269.3
```

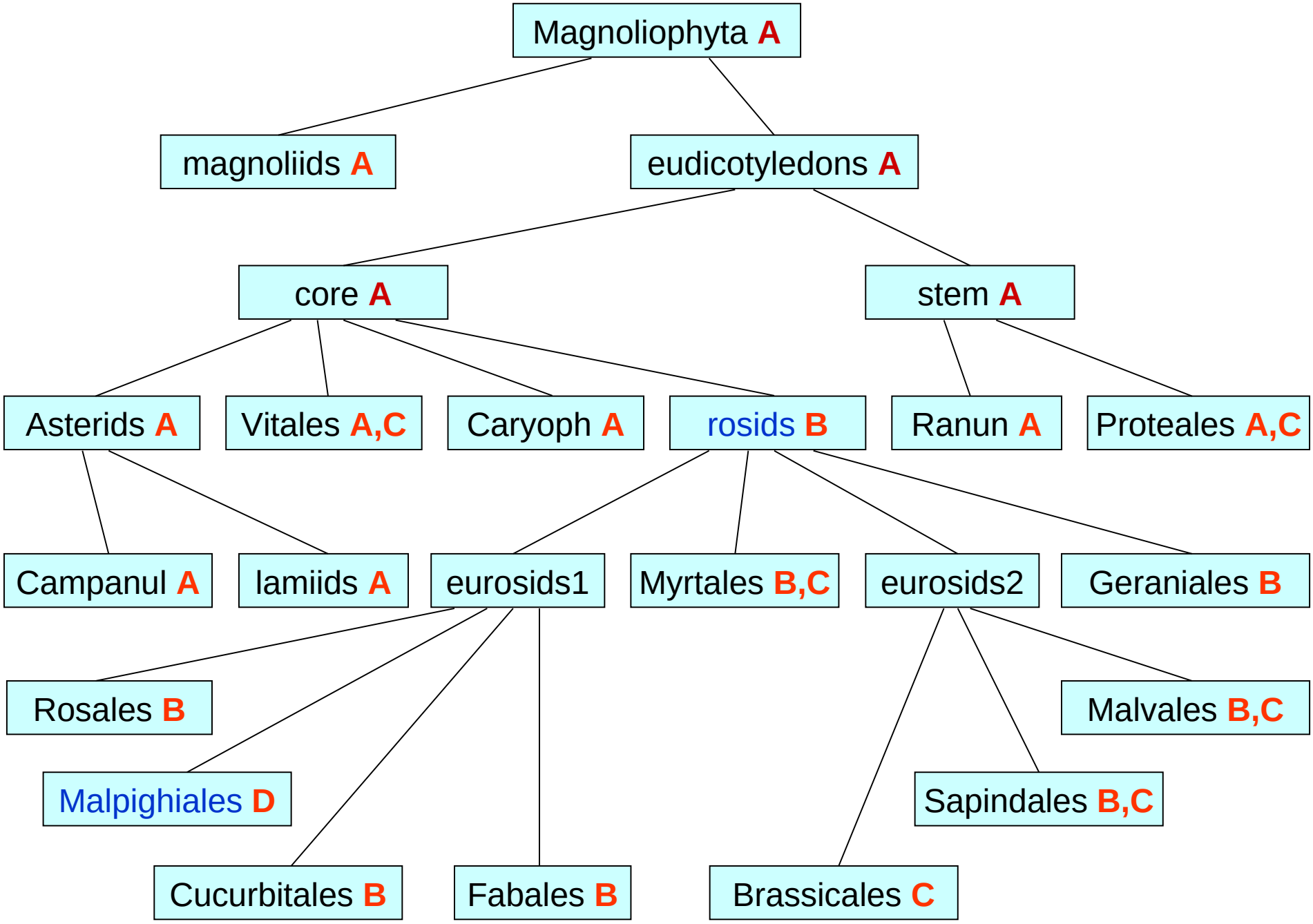
Поиск и эволюция сигнала другого типа («промотора»): некоторой комбинации слов с условиями на них и расстояния



На следующем слайде показан удивительно консервативный (=устойчивый при эволюции) прмотор (перед геном psbA в пластидах)

На слайде через один показан противопо-ложный случай: быстро эволюционирующий (меняющийся) промотор среди цветковых растений (перед геном *ndhF* в пластидах). Он имеет четыре варианта A, B, C, D, сменяю-щие друг друга. Сами эти промоторы найде-ны, но здесь не приведены.


TTGACATGGCT=ATATAAGTCATGTTATACT Arabidop
TTGACACGGG=CATATAAGGCATGTTATACT…ASpinacia
TTCACGATA==TATATAAGTCATACTATACT Cycas
TTGACATACA=GATATGTCTCATATTATACT Cryptomer
TTGACATTGAT=ACATGGATCATATTATACT Pinus
TTGACTTTAAT=AAACCATTTCTGTTATACT Welwitsch
TTGACACGGAT=AGGTTTTT=GTGATATGCT Adiantum
TTGACATCAAT=AGATAAGTTGTGTTATACT Angiopter
TTGACATATAT=GGAAAGATCATGTTATACT Psilotum
TTGACACAAA=AAGAAAGATTGTGTAATATT Huperzia
TTGACATAC=TAATGGGATATGTGTAATAAT Aneura
TTGACATAA=TCATATGTTATGTGTAATACT Marchantia
TTGACATAA=TAATACATTTTGTGTAATACT Physcomitr
TTGACATTT=TTATACTTTACATACTATAAT Chara
TTGACATTAGTTATACGT=TTGTGCAATACT Chaetospha
TTGACAGCT=TAAGGTTAAT=ATGTAATAAT Staurastr
TTGACAACAG=CATTAACTATCTGTAATAAT Zygnema
TTGACAAATA=AACATCATTT=TGGCATAAT Mesostig
TTGATTAATATAA=ATTAATTA=GTTATAAT Bigelowiel


Magnoliophyta A
magnoliids A
eudicotyledons A
core A
stem A
Asterids A
Vitales A,C
Caryoph A
rosids B
Ranun A
Proteales A,C
Campanul A
lamiids A
eurosids1
Myrtales B,C
eurosids2
Geraniales B
Rosales B
Malvales B,C
Malpighiales D
Sapindales B,C
Cucurbitales B
Fabales B
Brassicales C

Пример интересной темы для исследования – **связь** (PEP) **промоторов и предпочитаемых ими сигма-субъединиц**.

Например, нами показано, что промотор C предпочтительно связывает Sig4-субъединицу РНК-полимеразы. Аналогично для фаговых промоторов и полимераз.

Переходы, возможные в нашей модели регуляции, которая связана со спиралями:

(1) Правый конец ***y*** окна **сдвигается** на один нуклеотид вправо **или остается** на месте **или подается** сигнал **«T»**. Альтернатива: когда правый конец ***y*** доходит до начала гена, то **подается** сигнал **«A».**
При этом вторичная структура в окне формирует выбор между T или A;

(2) Левый конец ***x*** окна **сдвигается** на три нуклеотида вправо **или остается** на месте, что зависит от частоты ***c*** предшествующего считывания регулируемого гена;

(3) Вторичная структура **преобразуется** в окне, т.е. текущая вторичная структура ω трансформируется в новую структуру ω'.

В модели с предыдущего слайда ищется (выход алгоритма) зависимость ***p*(*c*)** **– частота наступления состояния «Т» (несчитывания гена)**, при каждом фиксированном значении частоты считывания («концентрации») ***c***.

При наличии такой регуляции график ***p*(*c*)** имеет вид, показанный на слайдах 24 и 25. При ее отсутствии график ***p*(*c*)** имеет вид почти постоянной функции или даже убывающей функции.

Что можно читать по этим темам:

1a) тип сигнала – «вторичная структура»:

[Lyubetsky, Pirogov, Rubanov, Seliverstov, 2007, Journal of Bioinformatics and Computational Biology, vol 5, no 1, p. 155-180],

1b) тип сигнала – «промотор»:

[Селиверстов, Лысенко, Любецкий, 2009, Физиология растений РАН, том 56, № 5; Seliverstov, Lyubetsky Молекулярная биология, представлена]

2) Модели эволюции этих регуляций, т.е. эволюции сигналов 1a и 1b:

[Любецкий, Жижина, Рубанов, 2008, Гиббсовский подход в задаче эволюции регуляторного сигнала экспрессии гена, ППИ, №4; Горбунов, Любецкий МолБио, представлена]

Статьи можно получить от авторов по адресу: gorbunov@iitp.ru

Наши биологические результаты (дает некоторый обзор, для слушателей не обязателен)

1. **Проведена реконструкция эволюционных событий молекулярного уровня:**

   построены деревья белков и согласующие их деревья видов, найдены события потенциальных горизонтальных переносов, потерь и дупликаций генов, случаи массовой дупликации генов в предковом геноме, статистические характеристики эволюционных событий по вершинам дерева видов и по таксономическим группам, сравнивались сценарии горизонтальных переносов против дупликаций и потерь генов. [In the book: Bioinformatics of Genome Regulation and Structure II. Springer Science & Business Media, Inc. 2005]

2. Предложены новые типы регуляции экспрессии генов:

2.1 Регуляция на уровне *трансляции*, опосредованная Т-боксом, например, гена *ileS*, кодирующего изолейцил-тРНК синтетазу, у Актинобактерий. [BMC Microbiology, 2005, 5:54; Молекулярная биология, 2005, 39(6)]

2.2 Регуляция на уровне *трансляции* посредством взаимодействия рибосомы, транслирующей лидерный пептид, и вторичной структуры РНК для гена *leuA*, кодирующего 2-изопропилмалатсинтазу, у Актинобактерий («LEU-элемент»). [BMC Microbiology, 2005, 5:54; Молекулярная биология, 2005, 39(6)]

2.3 Сложные типы классической аттенюаторной регуляции (когда антитерминатор не альтернативен терминатору), например, у лактобацилл перед геном *ilvD*: это – цепь спиралей или псевдоузел. [готовится к печати]

2.4 Аттенюаторная регуляция генов *cysK* синтеза цистеина у Актинобактерий, вовлекающая ро-белок для терминации транскрипции: рибосома, транслирующая лидерный пептид, перекрывает сайт связывания ро-белка. [BMC Microbiology, 2005, 5:54]

2.5 Регуляция гена *leuA* у альфа-протеобактерий, вовлекающая ген лидерного пептида и консервативный псевдоузел («LEU1-регуляция»). [готовится к печати]

2.6 Регуляция, опосредованная аномально длинной спиралью РНК, генов, кодирующих транспортёры двухвалентных катионов (*mntH*) и ферменты, зависимые от металлов (никель-зависимая глиоксалаза и др.), у бруцелл. Выясняется роль этой регуляции в выживании бруцеллы при незавершённом фагоцитозе (бруцеллез). [Биофизика, в печати]

2.7 Статистические данные о расположении длинных спиралей в геномах Актинобактерий относительно кодирующих областей: длинные спирали концентрируются в некодирующих областях вблизи 3'-концов высоко экспрессируемых генов (включая тРНК) или между сходящимися навстречу друг другу генами. Выясняется роль таких шпилек в снятии конформационного напряжения ДНК и при терминации транскрипции путем образования крест-шпилек на ДНК. [МолБиол, 2007, 41(4)]

3. Найдены новые случаи известных типов регуляции у бактерий:

3.1 Предсказана белок-ДНКовая регуляция на уровне транскрипции и также промоторы генов синтеза пролина у протеобактерий родов *Pseudomonas* и *Shewanella.* [Молекулярная биология, 2007, 41(3)]

3.2 Предсказано много случаев белок-ДНКовой репрессии/активации. В частности, охарактеризован GlpR-регулон (регуляция метаболизма глицерол-3-фосфата). [Молекулярная биология, 2003, 37(5) – совместно с М.С. и его сотрудниками].

3.3 Проведен широкомасштабный поиск регуляции на уровне транскрипции посредством Т-боксов. [Молекулярная биология, 2005, 39(6)]

3.4 Предсказана классическая аттенюаторная регуляция: (a) у протеобактерий (включая дельта-протеобактерии) и у видов из таксономических групп бацилл/клостридий и бактероидов [FEMS 2004], (b) у Актинобактерий [BMC Microbiology, 2005, 5:54]

3.5 Предсказана регуляция на уровне трансляции посредством тиаминового рибопереключателя для гена *ykoE*, кодирующего субъединицу ABC транспортёра: происходит перекрывание сайта связывания рибосомы иногда прямо черенком рибопереключателя, а иногда дополнительной спиралью РНК –

происходит быстрая смена этих механизмов регуляции у очень близких видов (показана эволюция этого механизма). [Информационные процессы, 2006, 6 (1)]

4. Белок-РНКовая регуляция в пластидах:

4.1 Корреляция сплайсинга с белок-РНКовой регуляцией трансляции в хлоропластах растений и водорослей. [Journal of Bioinformatics and Computational Biology, 2006, 4, 4, 783; Биофизика, 2006, 51, тематический выпуск 1]

4.2 Связь вторичной структуры РНК с редактированием инициирующего кодона в хлоропластах у мхов и папоротников. [Биофизика, 2006, 51, тематический выпуск 1]

4.3 Найдена высоко консервативная регуляция экспрессии генов *psaA*, *psbA* и *psbB* (вне связи со сплайсингом) [Journal of Bioinformatics and Computational Biology, 2006, 4(4)].

4.4 Найдена ортологичная консервативная регуляция гена *ycf24* на уровне трансляции в пластидах красных водорослей и паразитов из таксона Apicomplexa (*Eimeria tenella*, *Plasmodium* spp., *Toxoplasma gondii*). Более того, у *T. gondii* эта регуляция охватывает и много других генов, включая те, которые кодируют РНК-полимеразу:

этот ген кодирует белок SufB, необходимый для формирования железосероцентров.
Выясняется роль пластид в жизни токсоплазм на молекулярном уровне. [Мол. биология, в печати]

5. Промоторы бактериального типа в пластидах и соответствующие им сигма-факторы у растений и водорослей:

5.1 Изучена быстрая эволюция промоторов перед геном *ndhF*, чья транскрипция у Резушки Таля (*Arabidopsis thaliana*) существенно зависит от сигма-субъединицы Sig4. [Физиология растений, в печати].

5.2 Предсказано, что кодируемая в ядре сигма-субъединица Sig4 РНК-полимеразы бактериального типа существовала уже у предка высших двудольных растений и у него же имелся Sig4-зависимый промотор:

соответствующие кДНК *sig4* найдены по базе EST у винограда *Vitis vinifera* и двух видов апельсина *Citrus clementina* и *C. sinensis* (у апельсинов это псевдоген). Также известен псевдоген *sig4* у тополя *Populus trichocarpa*. А Sig4-зависимые промоторы предсказаны в хлоропластах у всех видов из таксона Eurosids II (включая крестоцветные, апельсин и хлопок), а также у нескольких далёких представителей двудольных: эвкалипта, винограда и платана.

5.3 Исследованы Sig3-зависимые промоторы перед геном *psbN* у семенных растений и показано общее! для всех однодольных растений значительное отличие области этого промотора от прочих цветковых растений.

5.4 Найдены высоко консервативные хлоропластные промоторы бактериального типа перед генами *rbcL*, *psaA*, *psbA*, *psbB*, *psbE* у большинства видов из Streptophyta.

Более того, промотор перед геном *psbA*, кодирующим белок D1 второй фотосистемы, одинаков у Streptophyta, включая рано отделившиеся роды *Mesostigma* и *Chlorocybus*, и у вторичного симбионта *Bigelowiella natans* из таксона Cercozoa.

5.5 Найдены промоторы перед геном *rps20* и близлежащие сайты связывания транскрипционного фактора (– ортолога NtcA) в хлоропластах красных и криптофитовых водорослей. При этом сайт для NtcA найден тогда и только тогда, когда дивиргентно располагается ген *glnB. У* цианобактерий оба белка NtcA и GlnB вовлечены в *регуляцию* генов метаболизма азота и их взаимная регуляция показана (в частности, NtcA активирует транскрипцию *glnB*).

На этом основании предсказана регуляция в хлоропластах по механизму конкуренции РНК-полимераз, транскрибирующих гены на противоположных цепях ДНК, причем также происходит активация транскрипции *glnB*.

6. Найдена общая белок-ДНКовая регуляция экспрессии ядерных генов, кодирующих рубредоксин и киназу, фосфорилирующую белки по тирозину,

у диатомовой водоросли *Thalassiosira pseudonana* и у паразитов родов *Theileria* и *Babesia*

Эти виды являются вторичными симбионтами и имеют пластиды с общим происхождением от красных водорослей. Однако их ядерные геномы сильно отличаются. Поэтому можно предполагать связь этой регуляции с пластидами. Интересно, что киназы обычно участвуют в регуляторных каскадах, передающих сигнал от некоторой мембраны, в частности, от пластиды.

Пластиды у диатомовых водорослей и паразитов Apicomplexa похожи, а ядерные геномы значительно различаются. С другой стороны, у криптофитовых водорослей рубредоксин кодируется в нуклеоморфе, т.е. непосредственно связан с пластидами. Поэтому можно предположить, что эти очень близкие регуляторные механизмы связаны с появлением пластид.