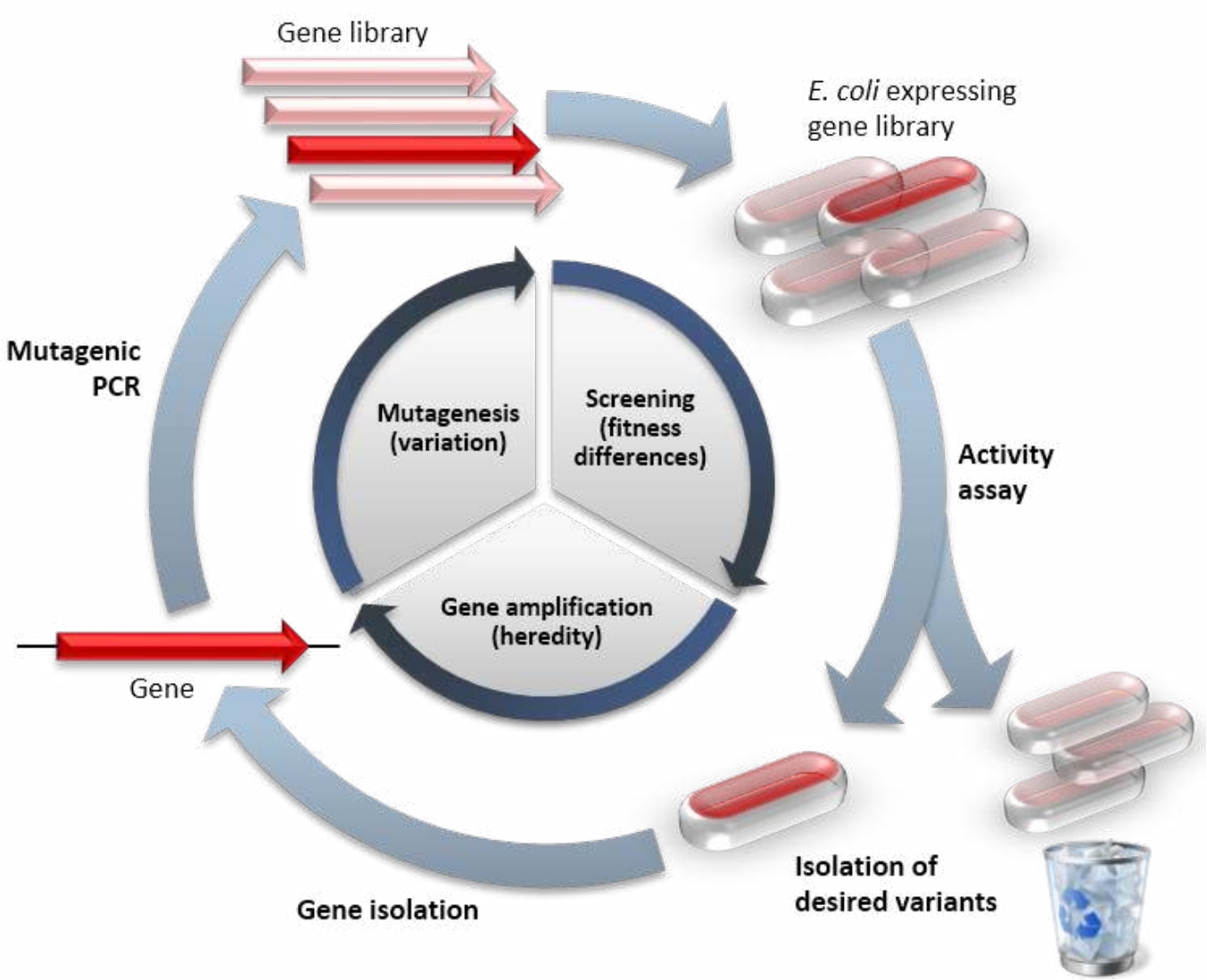
Биоинформатика в синтезе генетических конструкций

Павел Яковлев

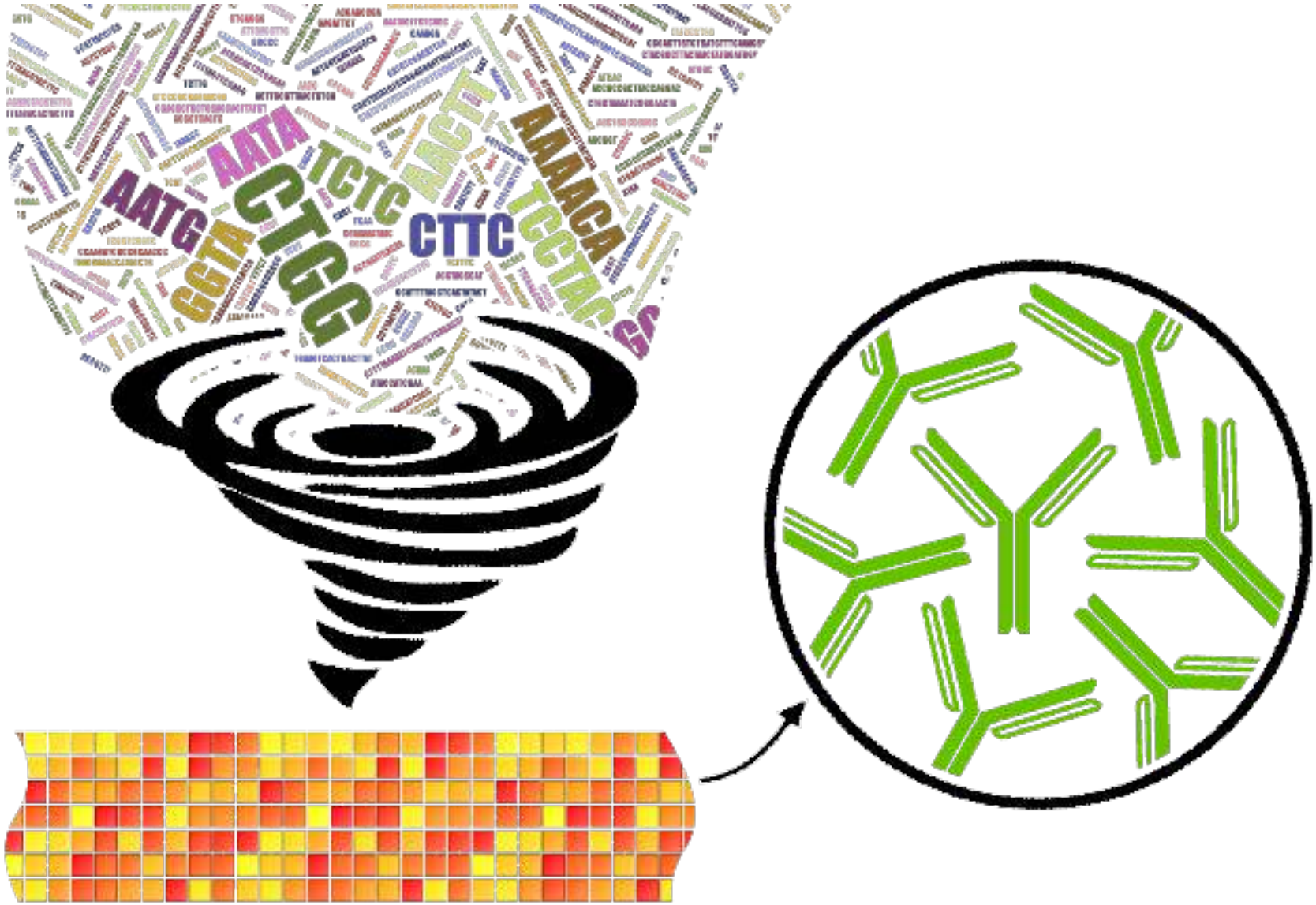
BIOCAD
Biopharmaceutical Company

23 июля 2014
Летняя школа по биоинформатике 2015

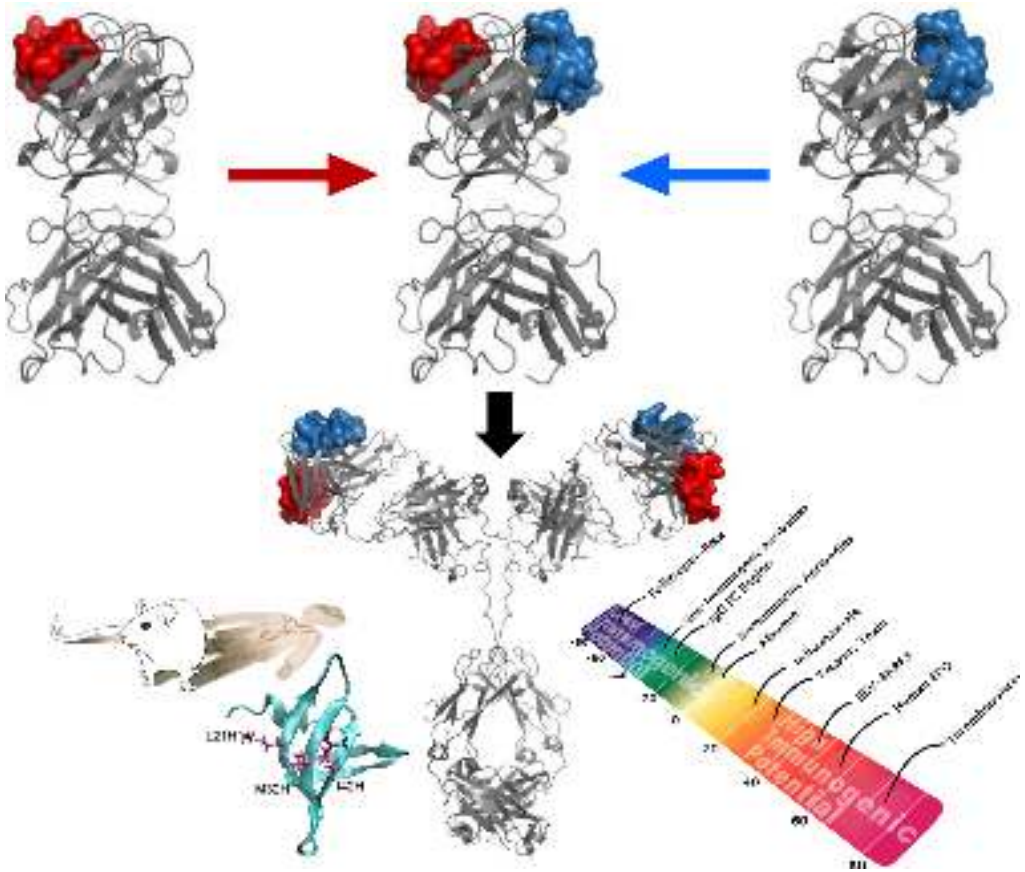
Направленная эволюция



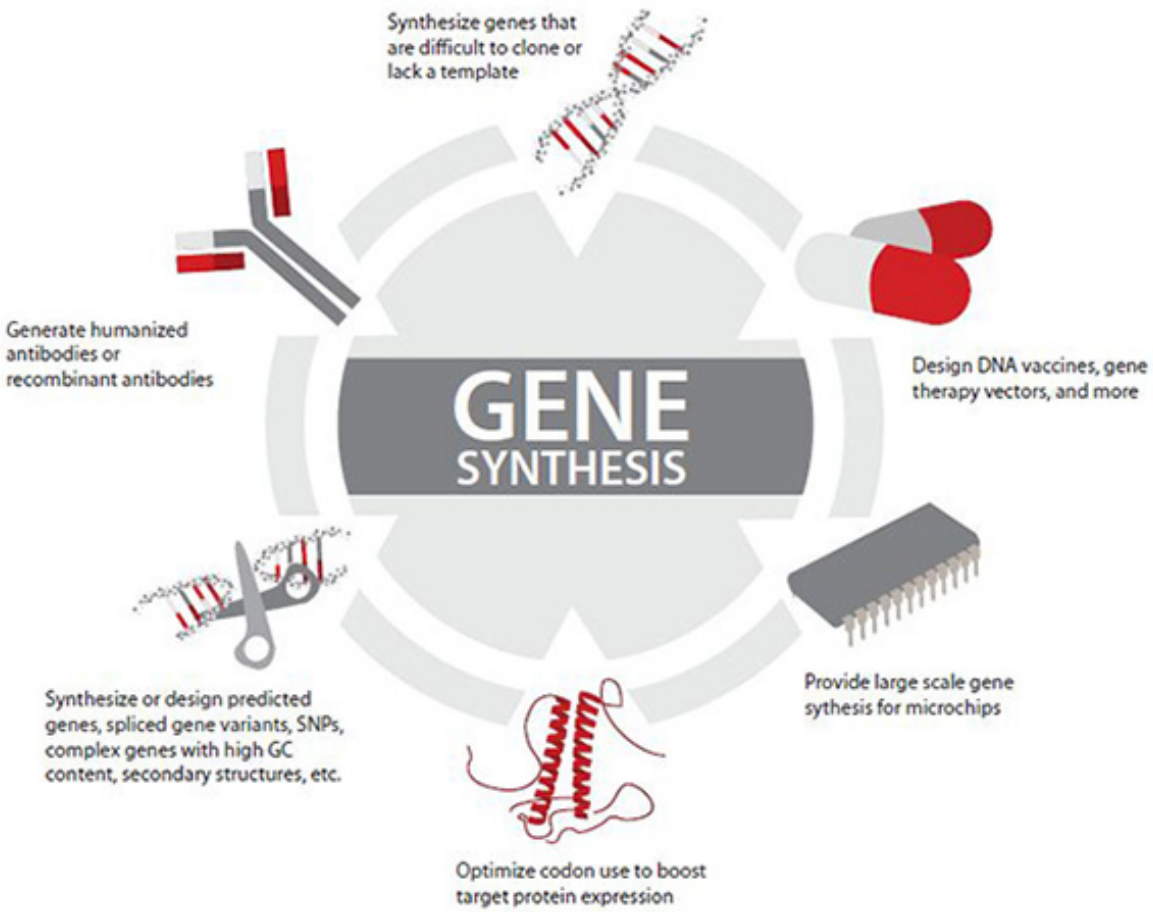
Секвенирование: *in vitro* $\rightarrow$ *in silico*



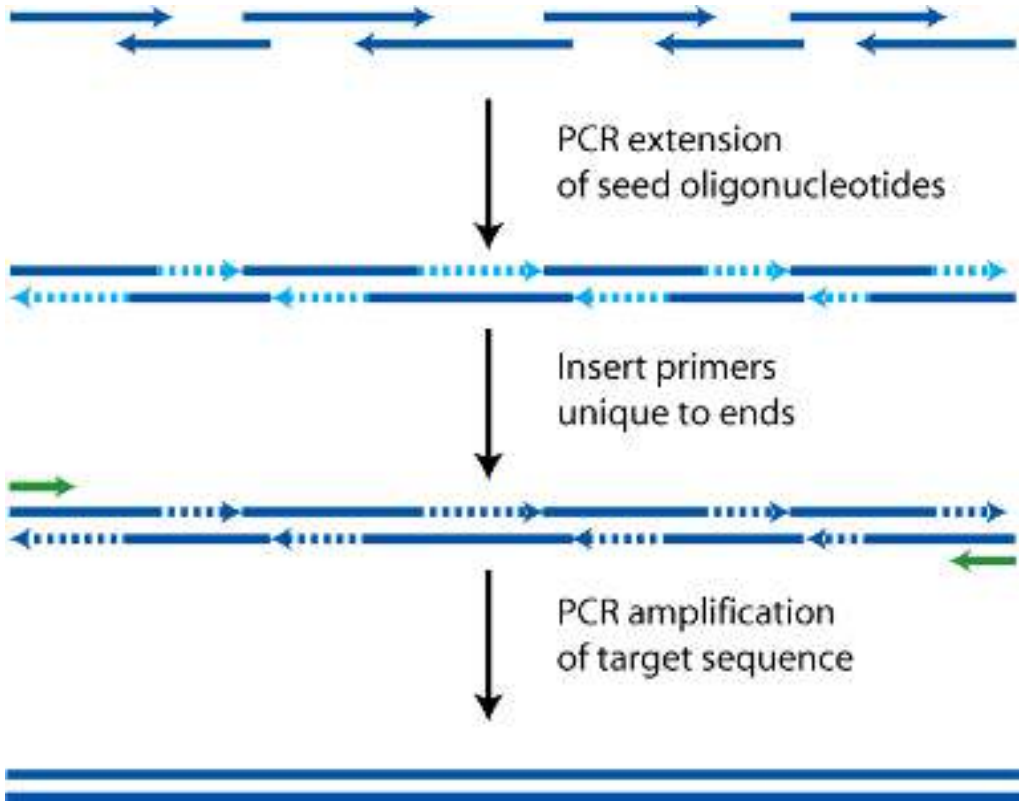
Дизайн белков



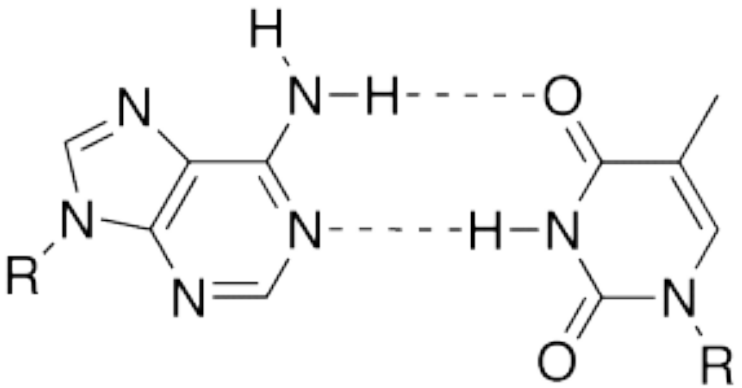
in silico $\stackrel{?}{\rightarrow}$ *in vitro*



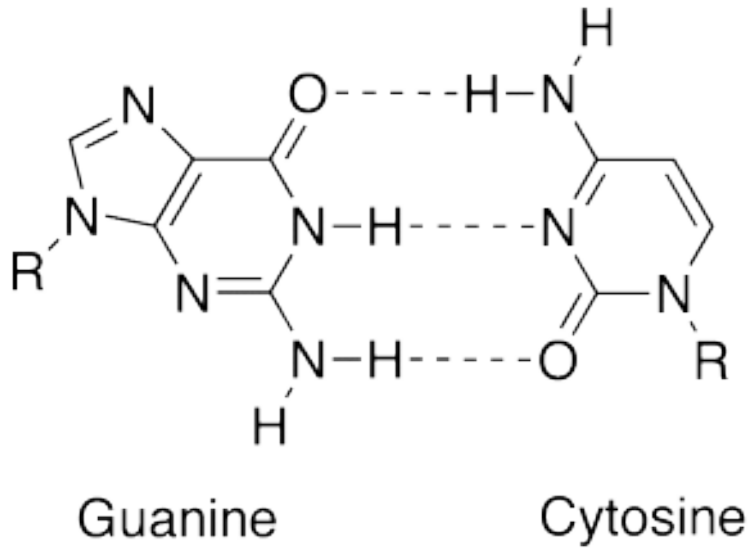
Polymerase Cycling Assembly



Комплементарность



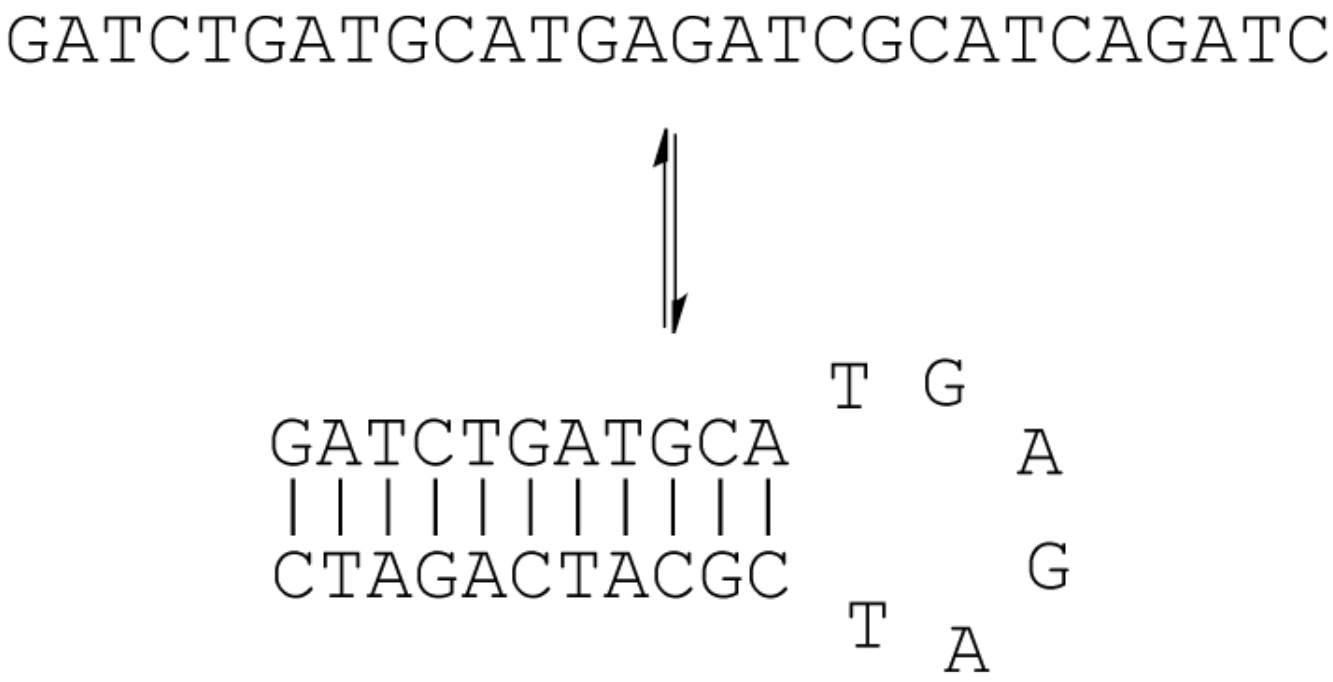
2 водородные связи



3 водородные связи

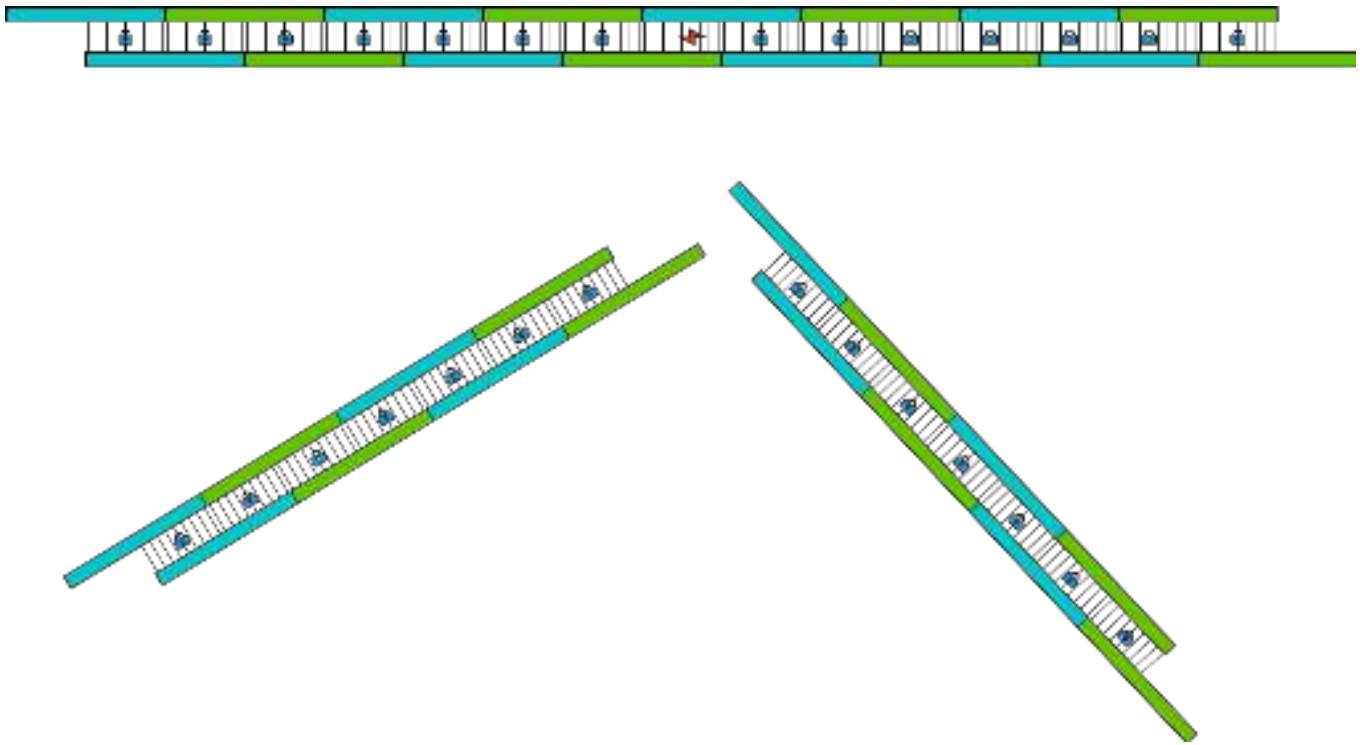
Проблемы с PCA I

- Шпильки



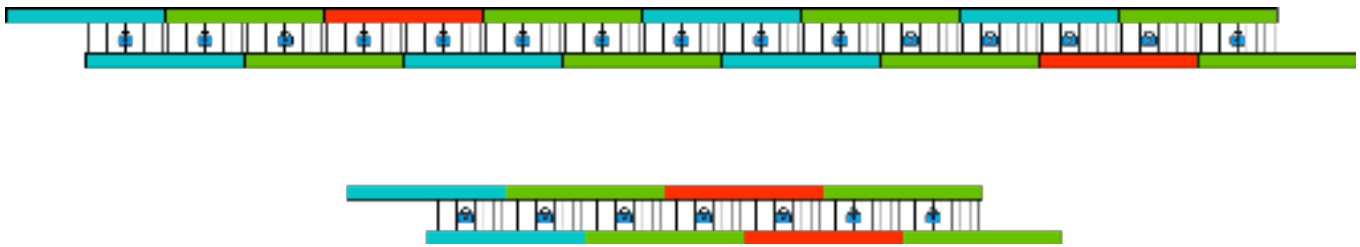
Проблемы с PCA II

- Слабые связи

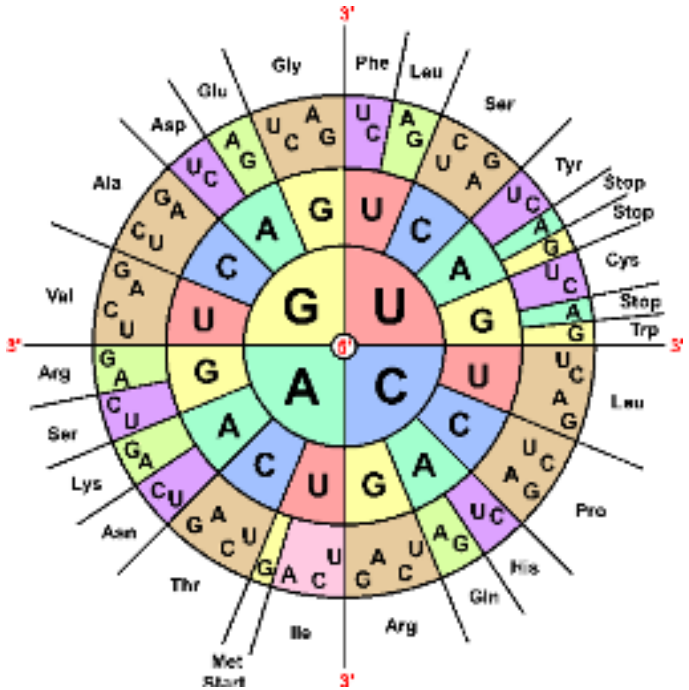


Проблемы с PCA III

- Кросс-активность



Переберем все варианты?



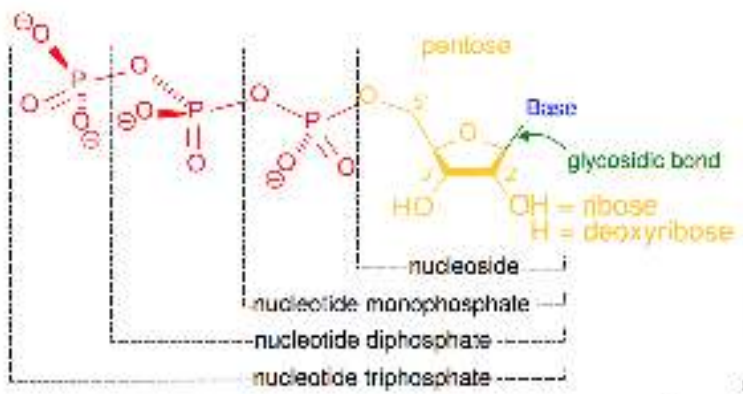
Количество вариантов кодирующих генов:

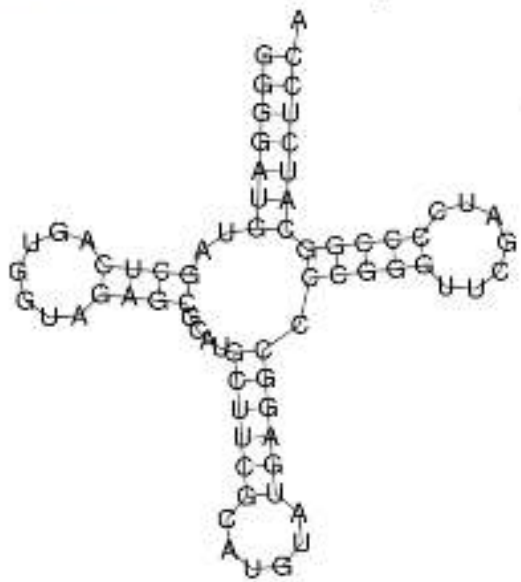
$$3^{100} \approx 5 \times 10^{47}$$

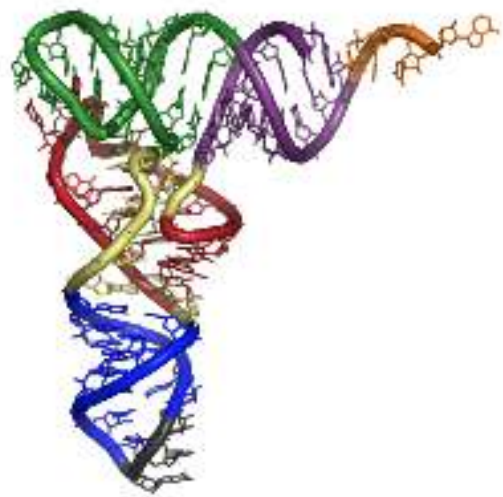
План по спасению

- Понять, как сворачиваются нуклеиновые кислоты.
- Понять, когда связи будут рваться, а когда нет.
- На базе полученных знаний построить оценочную функцию для группы олигов.
- Построить алгоритм, итеративно улучшающий имеющийся набор олигов.

Как сворачиваются ДНК и РНК?







Более формально I

Последовательностость РНК:

$$S \in \{A, C, G, U\}^*, \text{ с длиной } n = |S|.$$

Структура РНК:

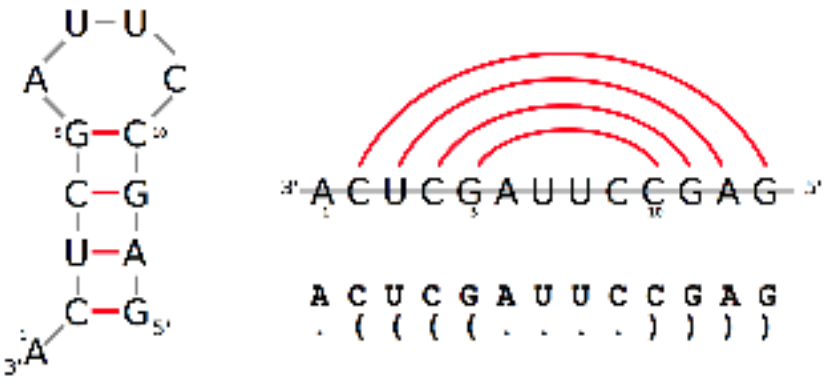
$$P \subseteq \{(i, j) | 1 \leq i \leq j \leq n, comp(S_i) = S_j\}.$$

Более формально II

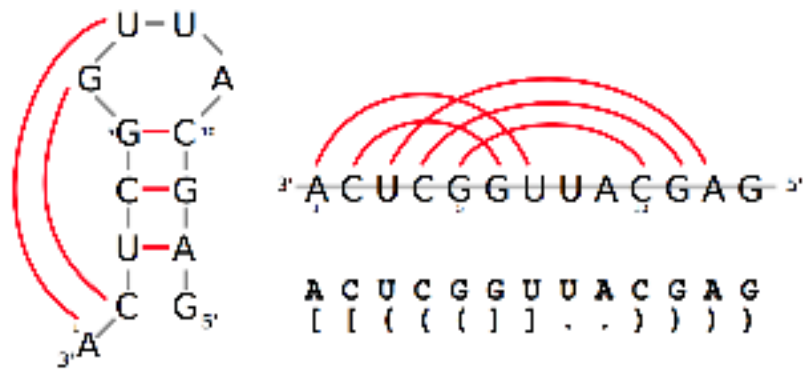
Будем называть две связи (i, j) и (i', j') пересекающимися (crossing), если

$$i < i' < j < j' \text{ или } i' < i < j < j'$$

Структура P является пересекающейся, если содержит хотя бы одну пару пересекающихся связей. Прочие структуры будем называть непересекающимися (non-crossing, nested).



$P = \{(2, 13), (3, 12), (4, 11), (5, 10)\}$



$P = \{(1, 7), (2, 6), (3, 12), (4, 11), (5, 10)\}$

Более формально III

При поиске структуры требуется определиться со следующими вопросами:

- Какую структуру считать корректной?
  - с макисмальным количеством связей;
  - с минимальной свободной энергией.
- Какой класс структур мы ищем?
  - crossing;
  - nested.
- Как мы хотим видеть ответ?
  - набор наиболее вероятных структур;
  - вероятность образования подструктур.

Простой выбор

Задача:

IN: Последовательность S

OUT: Одна nested структура P, максимизирующая количество связей.

Поиск crossing структур в общем случае – NP-hard.

Алгоритм Нуссинова I

Введем величину:

$$N_{i,j} = max\{|P| \mid P - \text{структура строки} S_{i,j}\}$$

В этом случае ответом на нашу задачу станет значение $N_{1,n}$, которое вы постараемся вычислить методом динамического программирования.

Алгоритм Нуссинова II

В случае $i = j$ мы будем иметь строки нулевой длины, соответственно:

$$N_{i,i} = 0$$

Аналогично, структуру не может образовать строка с отрицательной длиной:

$$N_{i,j} = 0, \;\; j < i$$

Как искать для всех прочих позиций?

- Требуется рассмотреть разлиные варианты образования структур между нуклеотидами S_i и S_j.

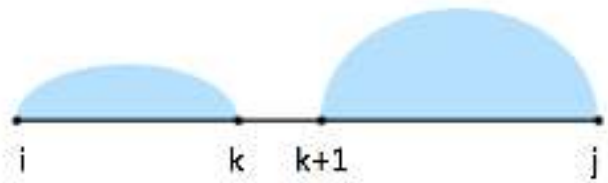
Алгоритм Нуссинова III

$N_{i+1,j}$

$N_{i+1,j-1} + \omega_{i,j}$

$N_{i,j-1}$

$$\max_{k:\, i<k<j} N_{i,k} + N_{k+1,j}$$



Алгоритм Нуссинова IV

Собираем:

$$N_{i,j} = 0,\ j \leq i$$

$$N_{i,j} = \max \begin{cases} S(i+1, j-1) + \omega_{i,j} \\ S(i+1, j) \\ S(i, j-1) \\ \max\limits_{k:\, i<k<j} N_{i,k} + N_{k+1,j} \end{cases}$$

Вариант 3 укладывается в вариант 4, а вариант 2 можно уложить, немного изменив лимиты на k.

Алгоритм Нуссинова V

$$N_{i,j} = \max \begin{cases} S(i+1, j-1) + \omega_{i,j} \\ \max\limits_{k:\, i \le k < j} N_{i,k} + N_{k+1,j} \end{cases}$$

G	C	A	C	G	A	C	G	
0								G
0	0							C
0	0	0						A
0	0	0	0					C
0	0	0	0	0				G
0	0	0	0	0	0			A
0	0	0	0	0	0	0		C
0	0	0	0	0	0	0	0	G

Алгоритм Нуссинова VI

$$N_{i,j} = \max \begin{cases} S(i+1, j-1) + \omega_{i,j} \\ \max\limits_{k:\, i \le k < j} N_{i,k} + N_{k+1,j} \end{cases}$$

G	C	A	C	G	A	C	G	
0	1							G
0	0	0						C
0	0	0	0					A
0	0	0	0	1				C
0	0	0	0	0	0			G
0	0	0	0	0	0	0		A
0	0	0	0	0	0	0	1	C
0	0	0	0	0	0	0	0	G

Алгоритм Нуссинова VII

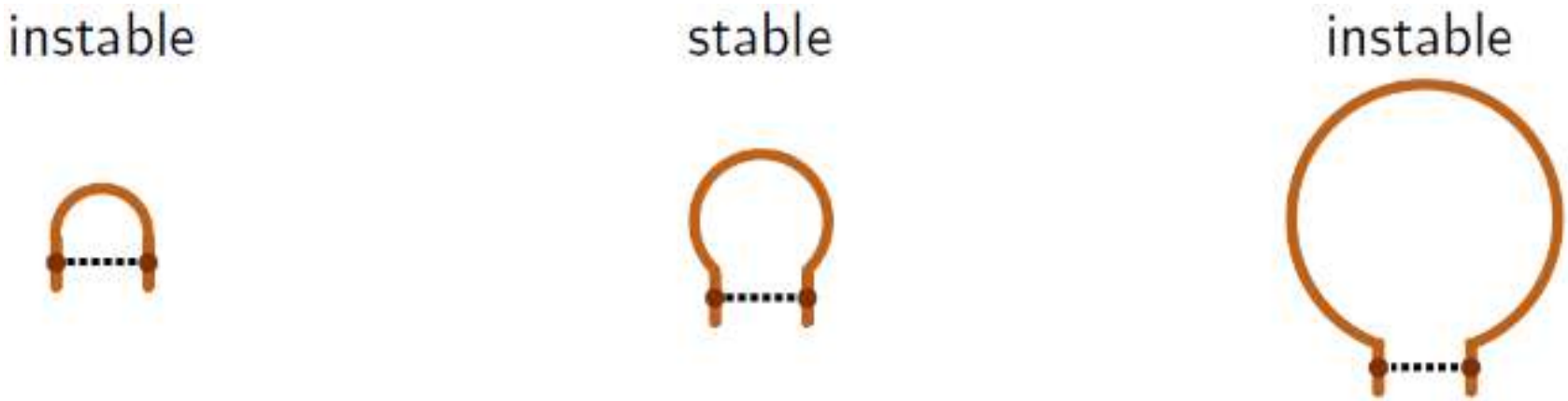
(	)	.	(	)	.	(	)	
G	C	A	C	G	A	C	G	
0	1	1	1	2	2	2	3	G
0	0	0	0	1	1	1	2	C
0	0	0	0	1	1	1	2	A
0	0	0	0	1	1	1	2	C
0	0	0	0	0	0	1	1	G
0	0	0	0	0	0	0	1	A
0	0	0	0	0	0	0	1	C
0	0	0	0	0	0	0	0	G

ACGA CGACG

Нахождение структуры: обратный проход от правого верхнего угла до диагонали.

Проблемы простого решения

- Максимизация взаимодействующих пар не отвечает реальности сворачивания.
- Взаимодействия влияют друг на друга, их нельзя рассматривать независимо.
- Существуют более и менее вероятные структуры.
- У РНК может быть более одной устойчивой структуры.



MFE-fonding I

Будем определять структуру исходя из значения *свободной энергии* высвобожденной в ходе формирования комплементарных пар.

Свободная энергия Гиббса – величина, показывающая изменение энергии в ходе химической реакции. Позволяет понять, возможно ли принципиально дальнейшее протекание химической реакции.

$$G = U - TS(+PV)$$

U – внутренняя энергия;
T – абсолютная температура;
S – энтропия.

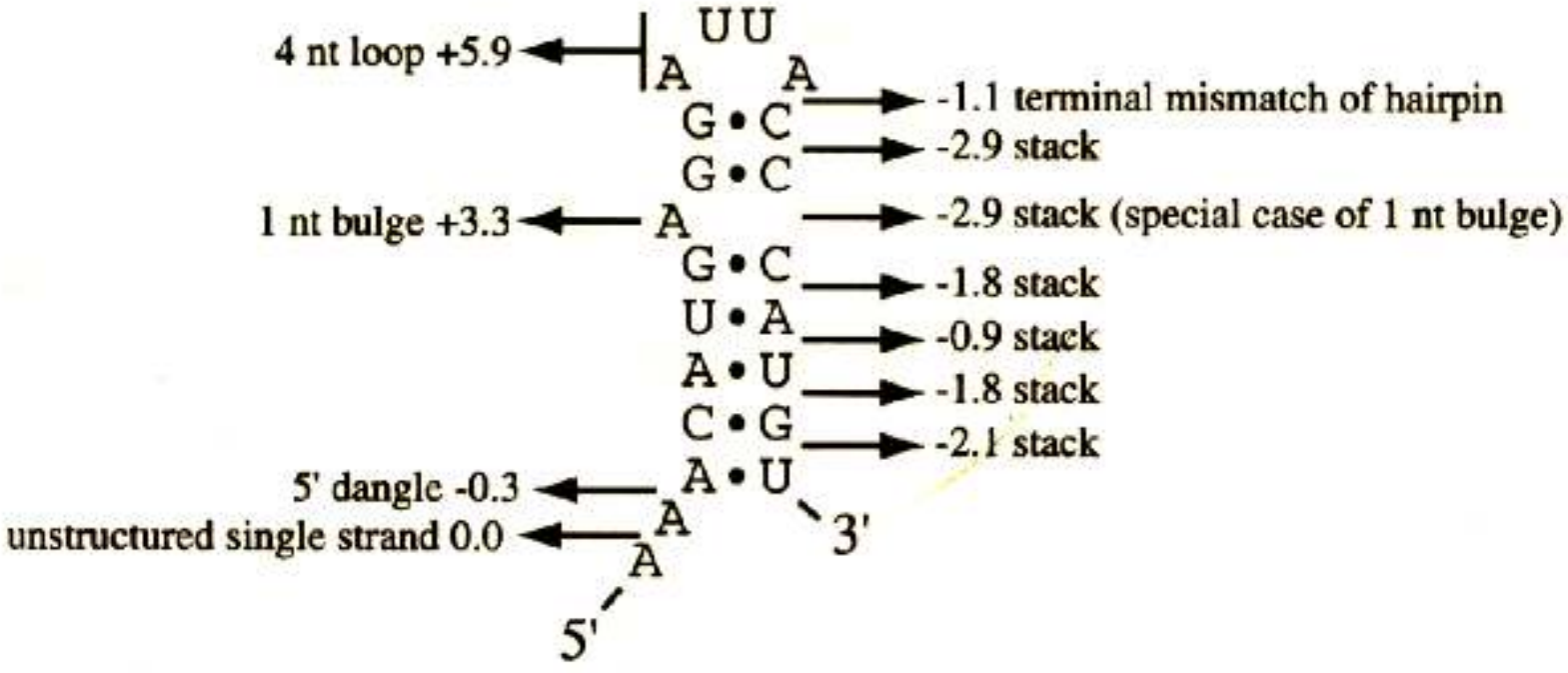
MFE-fonding II

- Будем предсказывать наиболее вероятную конформацию.
- Используем информауию об энергетическом статусе различных типов петель.
- Свободная энергия – аддитивная величина, а потому энергия структуры есть сумма энергий ее петель:

$$E(S) = \sum_{L \in S} E(L)$$

Алгоритм впервые предложен Цукером (Zuker) в 1981 году.

MFE-fonding III

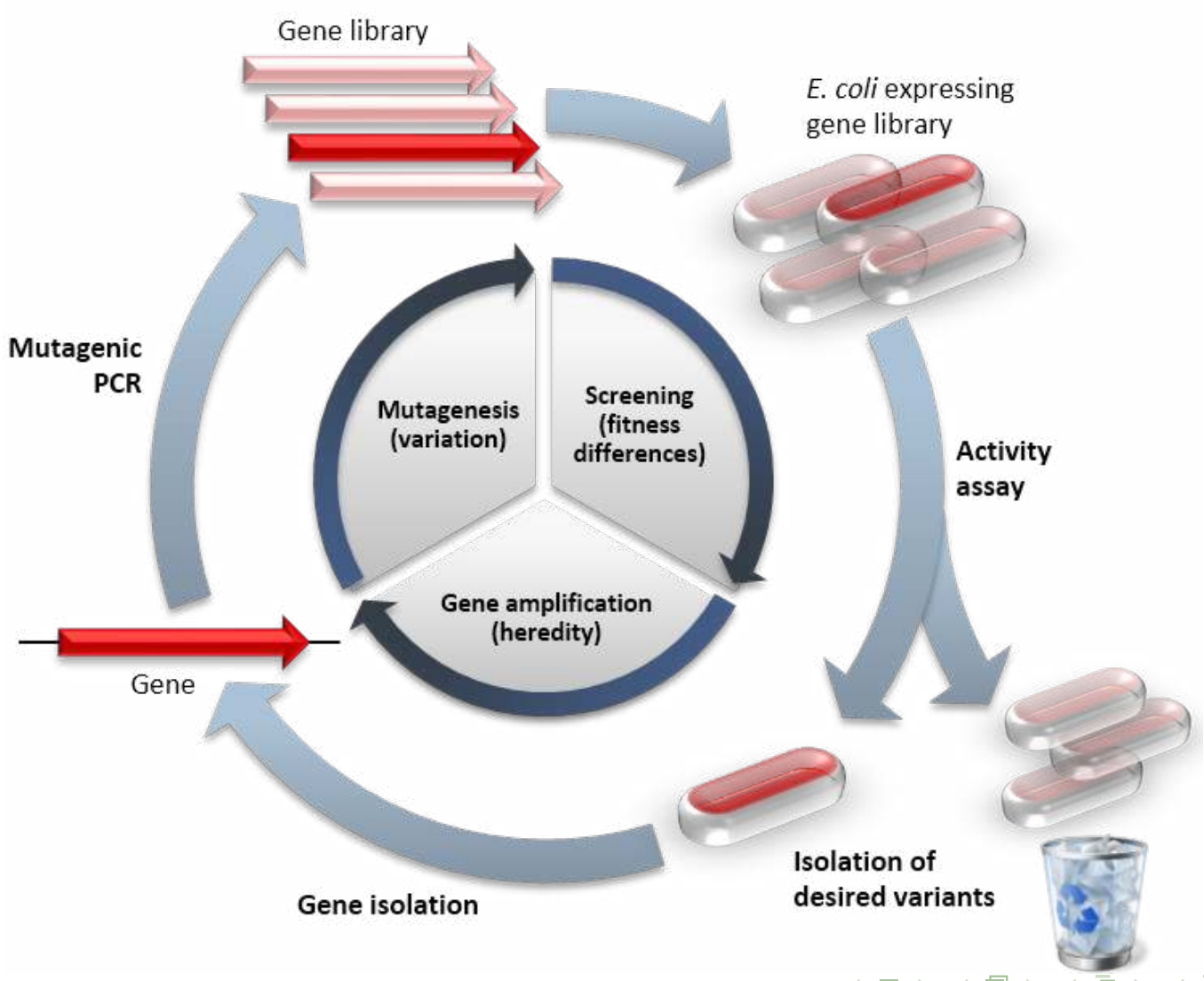


overall ΔG = -4.6 kcal/mol

К олигонуклеотидам

- Модифицированный алгоритм Цукера.
- Дает одно решение, но численно определяет MFE для заданной температуры, что позволяет оценивать качество олигов.
- Простая модификация позволяет использовать его же для оценки связывания нескольких олигонуклеотидов.

Genetic algorithm for genetics I



Genetic algorithm for genetics II

Вариант – улучшай худшего.

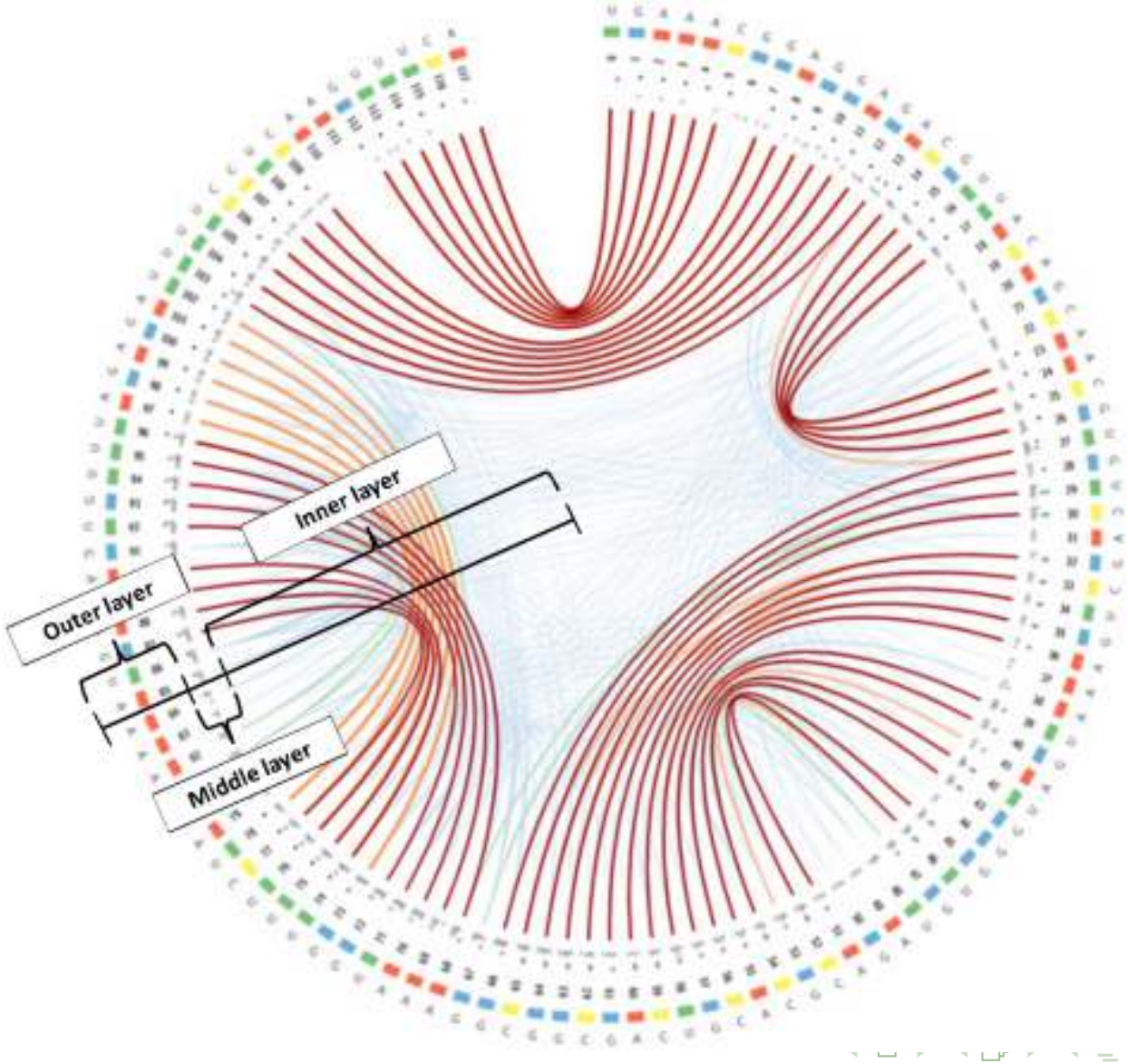
Алгоритм:

- Сгенерировать большой репертуар вариантов.
- Разбить варианты на олиги и оценить их.
- Выбрать лучший вариант.
- Исходя из трех компонент оценки выбрать область с олигами “худшего качества”.
- Методом Монте-Карло просемплирвоать область до достижения лучшего результата.
- Итеративно повторить необходимое количество раз со сменой области улучшения.

Features

- Работа с рандомизированными вариантами.
- Генерация библиотек.
- Поддержка встраивания в различные плазмиды.
- Поддержка переиспользования олигов и наличия константных фрагментов.
- Разбиение на фрагменты для сборки больших конструкций.
- Совместимость выхода с роботизированным оборудованием синтеза, подготовки и постановки ПЦР.

Full-mRNA optimization



Q&A

Спасибо за внимание!

yakovlev@biocad.ru

BIOCAD
Biopharmaceutical Company