

## Задачи анализа данных ДНК-микрочипов

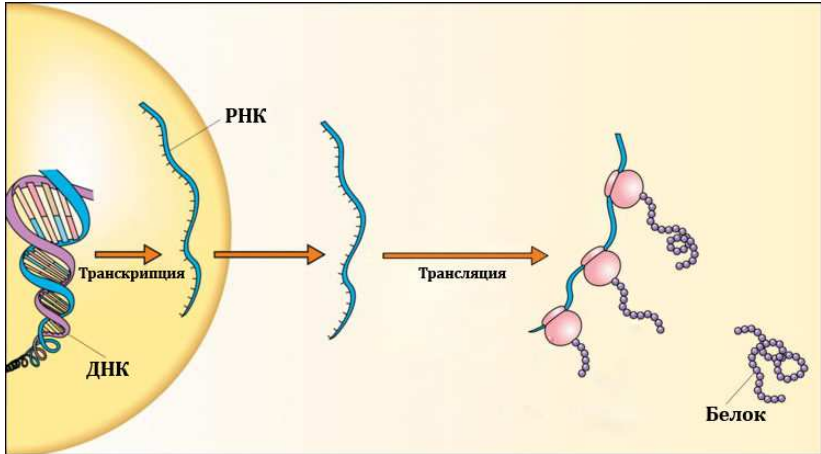
чл.-корр. РАН, д.б.н. Тоневицкий Александр Григорьевич  
чл.-корр. РАН, д.ф.-м.н. Рудаков Константин Владимирович  
д.ф.-м.н. Воронцов Константин Вячеславович

Семинар «Время, хаос и математические проблемы»  
МГУ • 19 октября 2011 г.

# Содержание

- 1 Введение
  - Базовые понятия молекулярной биологии
  - ДНК, РНК и гибридизация
  - Метаболические пути
- 2 ДНК-микрочип
  - Ход эксперимента и устройство ДНК-микрочипа
  - Сканирование ДНК-микрочипа
  - Стандартная методика оценивания экспрессии
- 3 Методы оценивания экспрессии
  - Простейшая модель данных ДНК-микрочипа
  - Модель, учитывающая эффект насыщения
  - Модель, учитывающая эффект кросс-гибридизации
  - Задача выявления активных метаболических путей

# Центральная догма молекулярной биологии



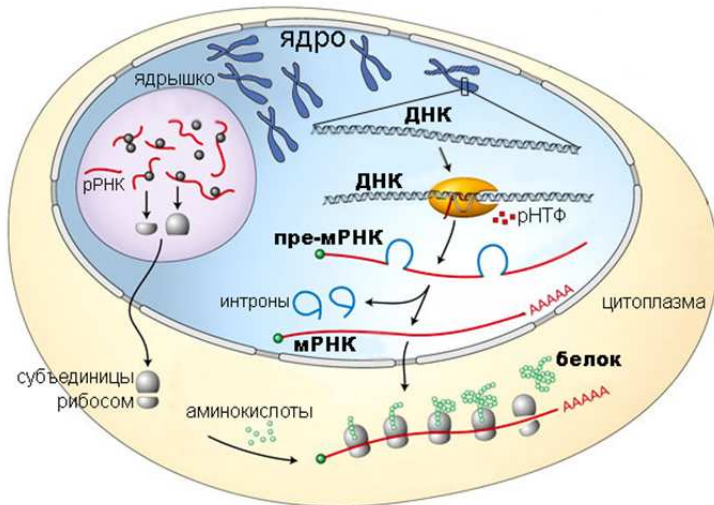
## Базовые понятия

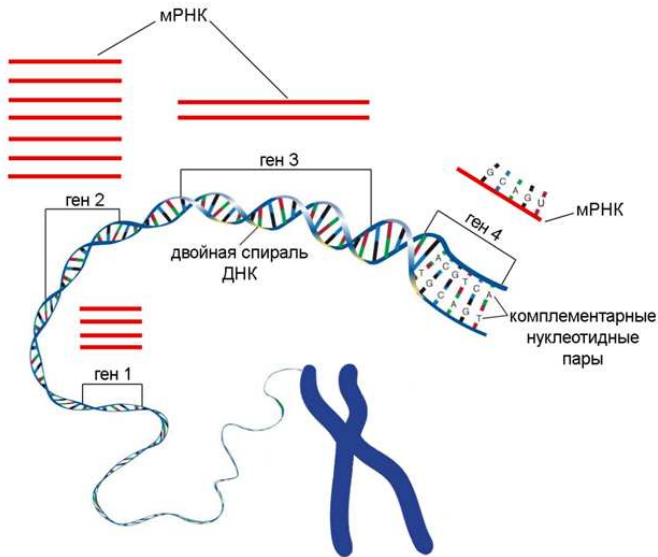
**ДНК** — молекула, содержащая информацию, необходимую для функционирования клетки.

**Ген** — участок ДНК, несущий какую-либо целостную функциональную информацию.

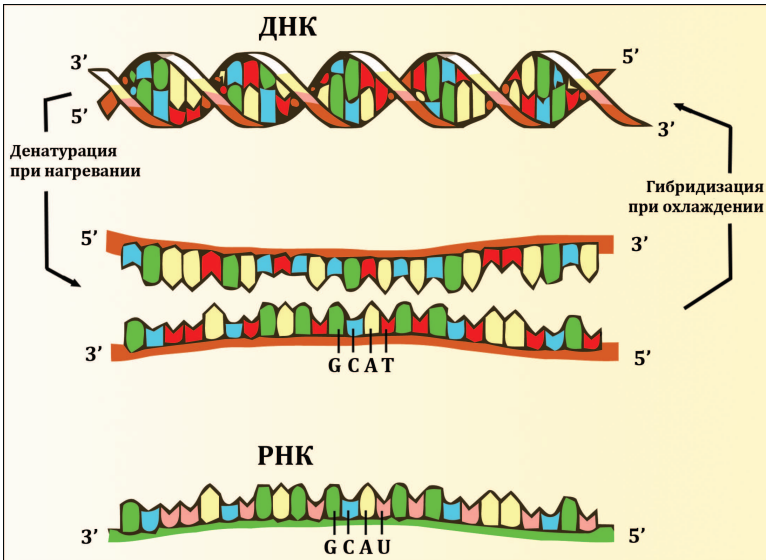
**РНК** — молекула-посредник, передающий информацию о гене структурам клетки, отвечающим за синтез белка.

Количество молекул РНК в клетке служит мерой активности гена (оценкой экспрессии).

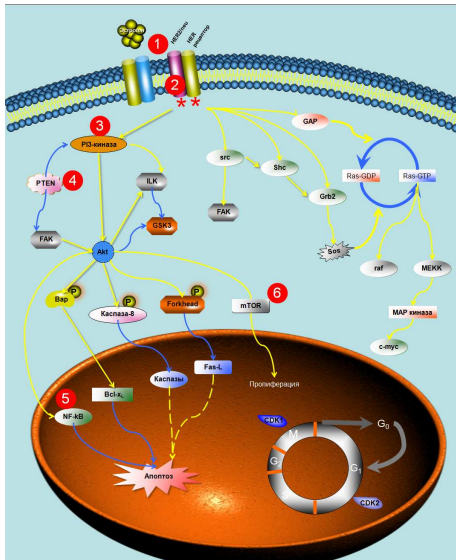




# ДНК и РНК

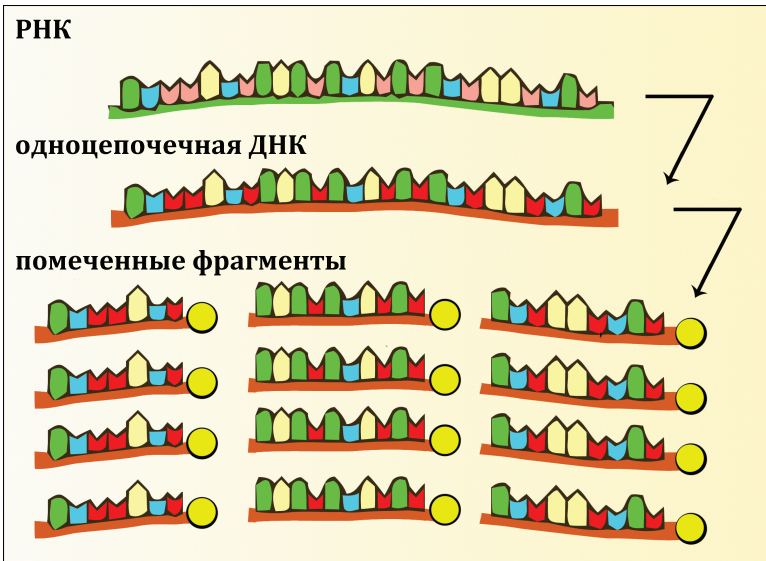


# Сети в клетке

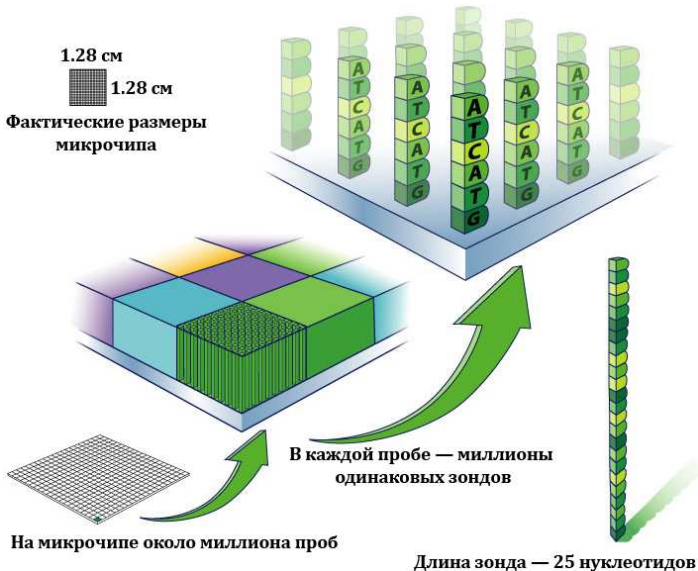




# Ход эксперимента

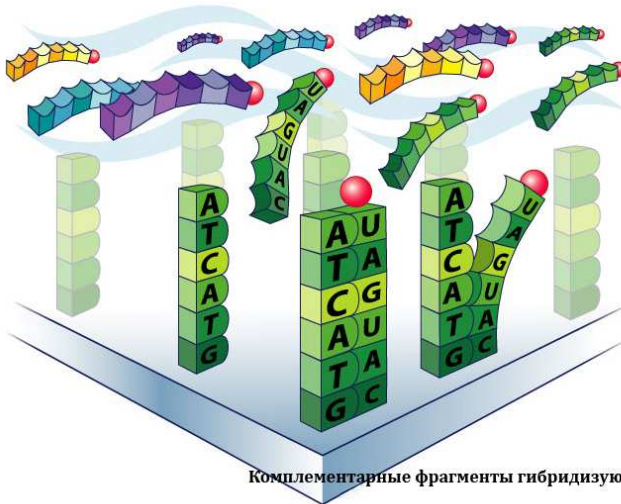


# ДНК-микрочип



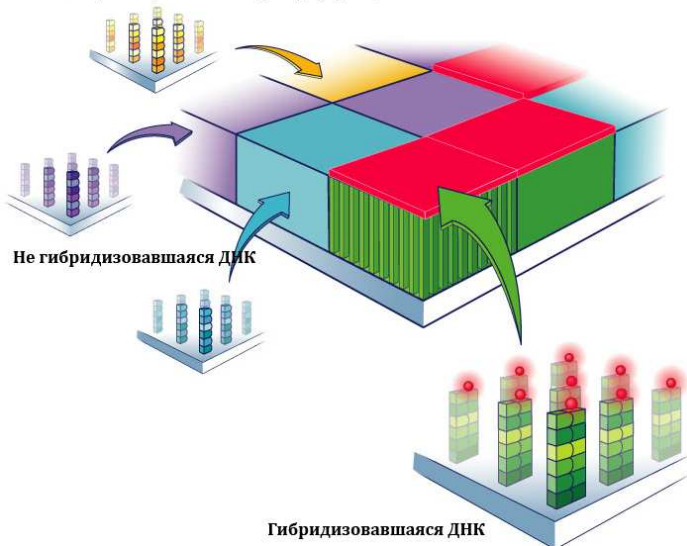
# Гибридизация

Помеченные фрагменты одноцепочечной ДНК наносятся на чип

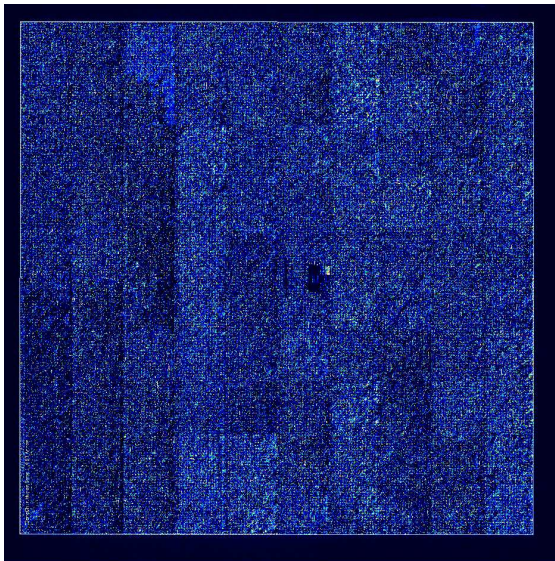


# Сканирование

При облучении лазером флуоресцентные метки светятся



# Результат сканирования

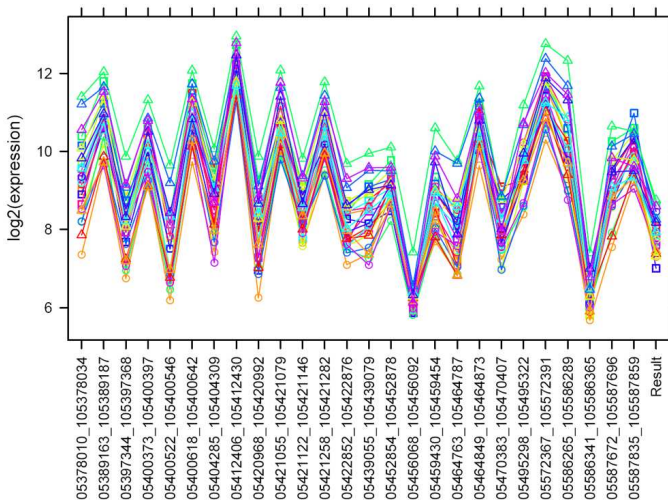


# Стандартная методика получения оценок экспрессии

- 1 Изображение со сканера оцифровывается, получается вектор значений интенсивности флуоресценции проб.
- 2 Проводится предобработка интенсивностей:
  - фоновая поправка;
  - нормализация.
- 3 Значения предобработанных интенсивностей всех проб каждого гена **усредняются** (median polish), давая оценку экспрессии.

# Пример получения оценки экспрессии

Интенсивности свечения проб одного гена для разных чипов:



# Модель №1, учитывающая степени сродства проб с геном

## Известные данные:

$I_p^k$  — интенсивность свечения пробы  $p$  на микрочипе  $k$ ;

$g(p)$  — номер гена, для которого проба  $p$  специфична  
(определён конструкцией микрочипа).

## Неизвестные параметры:

$C_g^k$  — концентрация РНК гена  $g$  на микрочипе  $k$ ;

$a_p$  — коэффициент сродства (affinity) пробы  $p$  гену  $g(p)$ ;

$N^k, B^k$  — нормировочные константы.

$$I_p^k = N^k a_p C_{g(p)}^k + B^k.$$

● ограничений = проб × чипов.  
861 493

● неизвестных = проб + (генов + 2) × чипов.  
861 493      28 869

Необходимо иметь хотя бы два микрочипа.



# Две стадии анализа данных ДНК-микрочипов

- 1 **Калибровка** параметров модели  $a_p$  по базе микрочипов

$$\sum_k \sum_p (I_p^k - N^k a_p C_{g(p)}^k - B^k)^2 \rightarrow \min_{\{a_p\}, \{C_g^k\}},$$

при ограничениях  $a_p \geq 0$ ,  $C_g^k \geq 0$  и нормировке  $\sum_g C_g^k = 1$ .

Данная задача декомпозируется по пробам.

- 2 **Восстановление концентраций**  $C_g^k$  для нового микрочипа

$$\sum_p (I_p - N a_p C_{g(p)} - B)^2 \rightarrow \min_{\{C_g^k\}},$$

при известных  $\{a_p\}$  и ограничениях  $C_g^k \geq 0$ ,  $\sum_g C_g^k = 1$ .

**Замечание.** Вместо функции потерь  $(I - \hat{I})^2$  могут использоваться  $|I - \hat{I}|$ ,  $|\log I - \log \hat{I}|$ , в общем случае —  $\mathcal{L}(I, \hat{I})$ .

## Модель №2 — уточнение

Одна проба может быть специфична нескольким генам.

$S(p)$  — множество генов, для которых проба  $p$  специфична.

$$I_p^k = N^k a_p \sum_{g \in S(p)} C_g^k + B^k.$$

$ S(p) $	проб	$ S(p) $	проб	$ S(p) $	проб	$ S(p) $	проб
1	721605	11	14	22	8	33	6
2	15352	12	872	23	2	34	1
3	1783	13	17	24	124	35	43
4	14416	14	83	25	569	36	423
5	222	15	206	26	3	38	3
6	2310	16	1170	27	6	39	8
7	44	17	1	28	66	40	53
8	634	18	143	30	134	42	176
9	5171	20	459	31	1	...	...
10	240	21	53	32	55	2220	1

# Модель №3, учитывающая эффект насыщения

Одна из простейших моделей насыщения — кривая Ленгмюра:

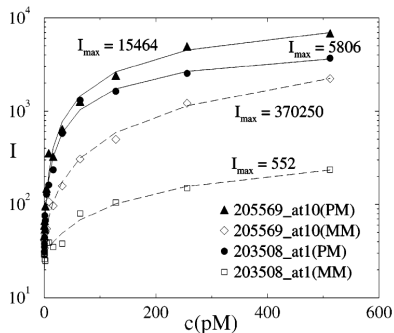
$$\sigma(C; a, b) = \frac{aC}{1 + bC}.$$

Модель с учётом насыщения:

$$I_p^k = N^k \sigma\left(\sum_{g \in S(p)} C_g^k; a_p, b_p\right) + B^k,$$

$a_p \geq 0$  — коэффициенты сродства,

$b_p = \frac{a_p}{I_{\max}}$ , где  $I_{\max}$  — уровень горизонтальной асимптоты.

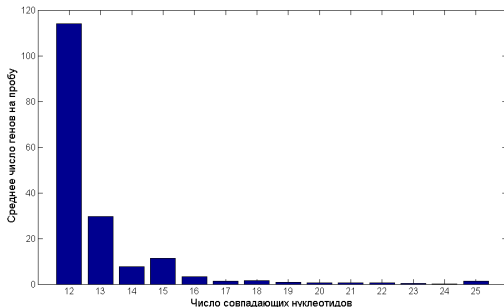


# Эффект кросс-гибридизации

Свечение пробы  $p$  может быть вызвано не только генами  $g \in S(p)$ , но и другими генами, для которых проба  $p$  НЕ специфична.

Проба 432:309		C	T	G	C	C	A	C	A	T	T	G	C	T	G	A	G	G	C	T	C	A	G	A	G	C	
Ген GRIA1	...	G	A	C	G	G	T	G	T	A	A	C	G	A	C	T	C	C	G	A	G	T	C	T	C	G	...
Ген GRIA3	...	G	A	C	G	G	T	G	T	A	A	C	G	A	G	T	C	C	G	A	G	T	C	T	C	G	...
Ген SNRPN	...	G	A	C	G	G	T	G	T	G	A	C	G	A	C	T	C	C	T	A	G	T	C	C	A	C	...
Ген DNAJC22	...	G	A	C	G	G	T	G	T	A	T	C	G	A	C	T	C	C	A	C	C	C	A	G	A	T	...

Распределение среднего числа комплементарных генов:



# Модель №4, учитывающая эффект кросс-гибридизации

**Гипотеза 1.** Свечение пробы  $p$  вызвано не только генами  $g \in S(p)$ , но и другими генами, для которых проба  $p$  НЕ специфична.

$\Gamma(p)$  — множество генов, которые могут гибридизироваться на пробе  $p$ ,  $\Gamma(p) \cap S(p) = \emptyset$ ,  $|\Gamma(p)| \lesssim 150$ .

$$I_p^k = N^k \sum_{g \in \Gamma(p)} a_{pg} C_g^k + B^k,$$

$a_{pg}$  — неизвестные коэффициенты сродства, матрица  $\|a_{pg}\|$  сильно разрежена.

**Гипотеза 2.** При неспецифической гибридизации насыщения нет:

$$I_p^k = N^k \left( \sigma \left( \sum_{g \in S(p)} C_g^k; a_p, b_p \right) + \sum_{g \in \Gamma(p)} a_{pg} C_g^k \right) + B^k.$$

## Открытая проблема

Оптимизационная задача на стадии калибровки:

$$\sum_k \sum_p \mathcal{L} \left( I_p^k, N^k \left( \sigma \left( \sum_{g \in S(p)} C_g^k; a_p, b_p \right) + \sum_{g \in \Gamma(p)} a_{pg} C_g^k \right) + B^k \right) \rightarrow \min_{\{a_p, b_p, a_{pg}\}, \{C_g^k\}},$$

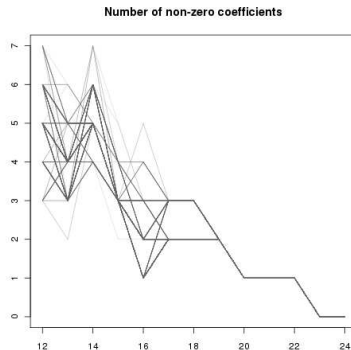
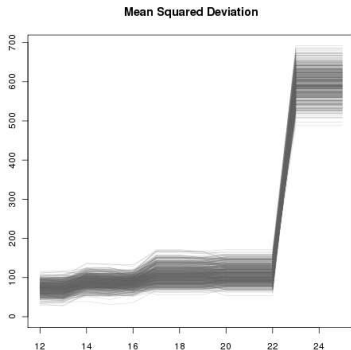
- ограничений = проб  $\times$  чипов.  
861 493
- неизвестных = проб  $(|\overline{\Gamma(p)}| + 2) + (\text{генов} + 2) \times$  чипов.  
861 493 28 869

## С чем связаны надежды:

- Достаточно нескольких сотен микрочипов (в базе более 3000).
- Задача «почти декомпозируется» по пробам.
- Стандартные методы, не учитывающие кросс-гибридизацию, дают неплохое начальное приближение.
- Технологии параллельных вычислений.

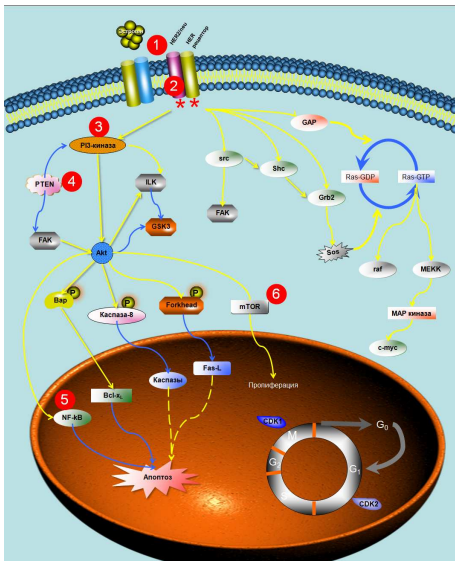
# Эксперимент по идентификации параметров $a_{pg}$

При уменьшении числа совпадающих нуклеотидов уменьшается невязка модели, увеличивается число ненулевых коэффициентов  $a_{pg}$ :



Планируемый следующий шаг — контроль переобучения по большой выборке чипов.

# Сети в клетке





# Задача следующего уровня

Выявление активных метаболических путей.

**Дано:**

1. Граф  $\langle V, E \rangle$ ,

$V$  — множество генов ( $\Leftrightarrow$  белков),  $V \approx 3 \cdot 10^4$ ,

$(x, y) \in E$  — белки  $x$  и  $y$  выполняют совместную функцию.

2. Множество метаболических путей

$$\mathcal{P} = \{P = (x_1, \dots, x_n) : (x_i, x_{i+1}) \in E\},$$

**Найти** по данным микрочипа ДНК:

Подмножество путей  $P = (x_1, \dots, x_n)$ , в которых все гены  $x_i \in P$  экспрессированы.

**Основная проблема:**

Граф  $\langle V, E \rangle$  и множество путей  $\mathcal{P}$  известны неполно и неточно.

# Спасибо за внимание!

чл.-корр. РАН, д.б.н. Тоневицкий Александр Григорьевич

д.ф.-м.н. Воронцов Константин Вячеславович

[voron@forecsys.ru](mailto:voron@forecsys.ru)

Презентация доступна на вики [www.MachineLearning.ru](http://www.MachineLearning.ru), страница  
<http://www.MachineLearning.ru/wiki/index.php?title=Участник:Vokov>