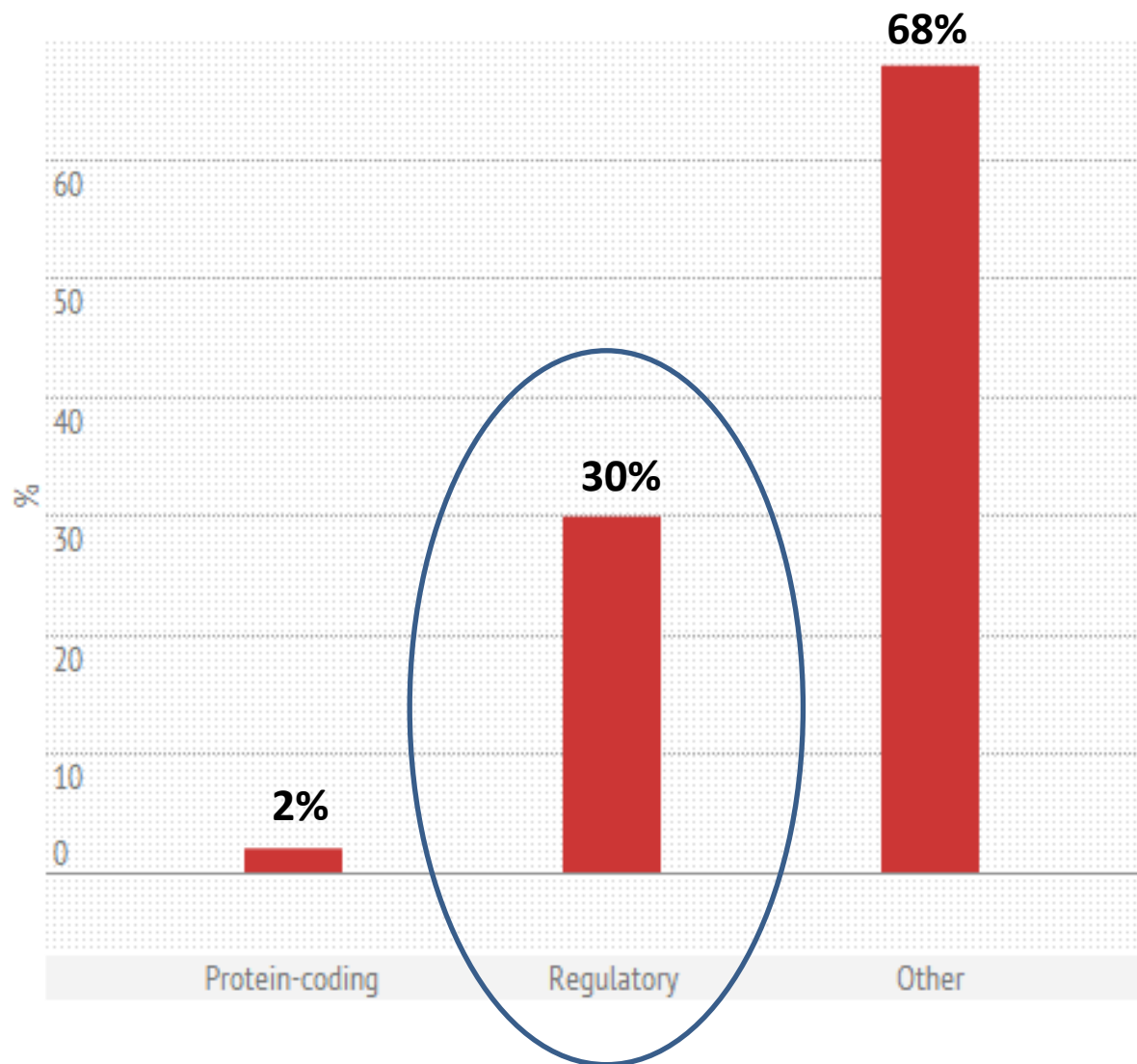


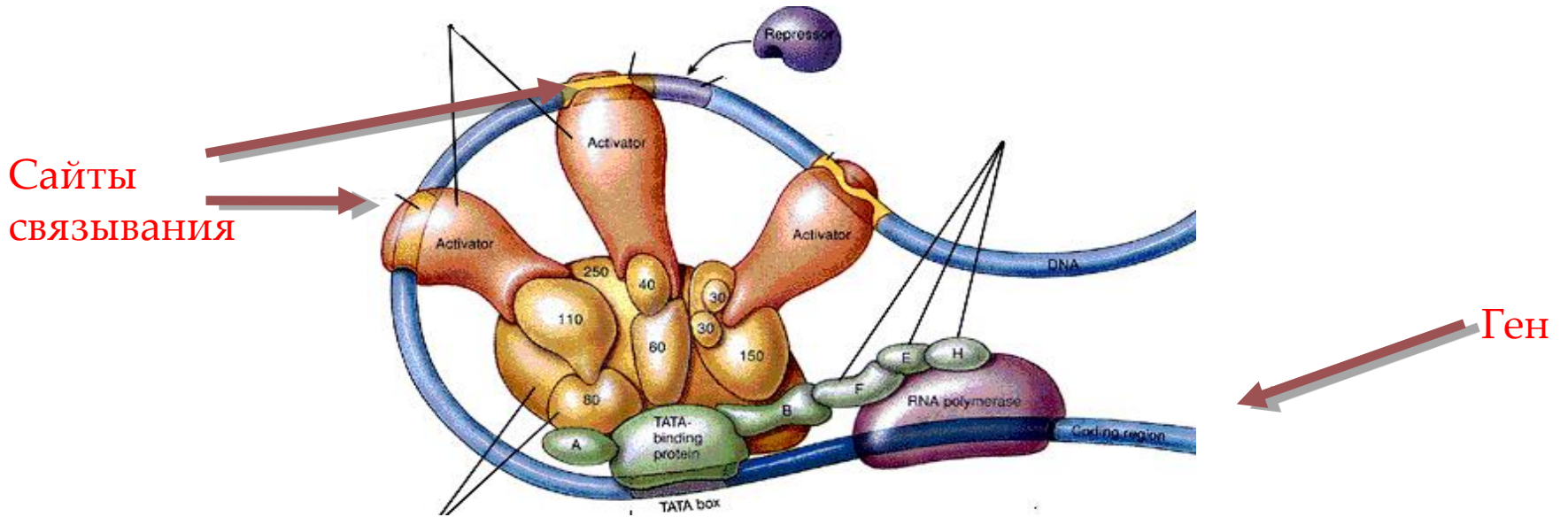
РЕАЛИЗАЦИЯ СРЕДСТВ  
ИЕРАРХИЧЕСКОГО  
АНАЛИЗА РЕГУЛЯТОРНЫХ  
РАЙОНОВ ГЕНОВ ДЛЯ  
ИНТЕГРИРОВАННОЙ  
СИСТЕМЫ EXPERT  
DISCOVERY И UGENE



*Юрий Васькин*

# Геном человека



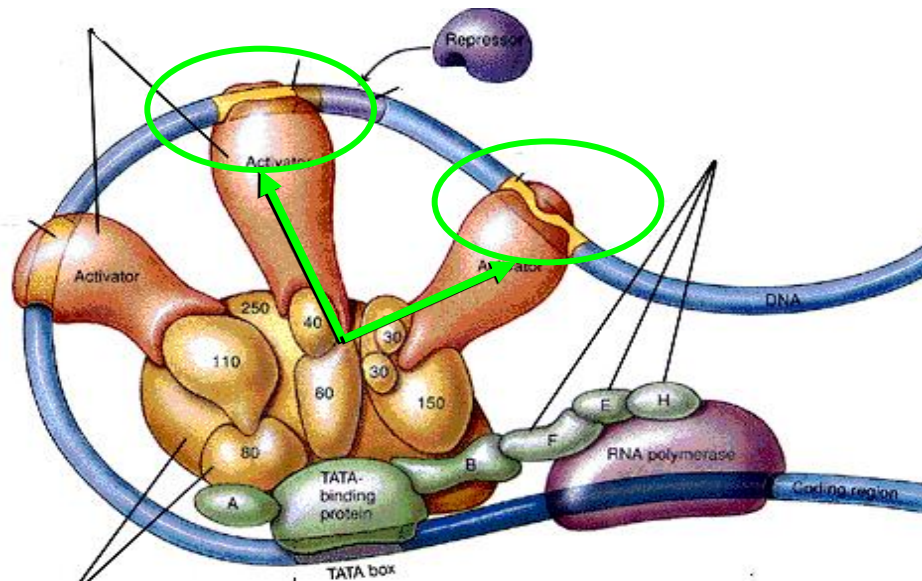


Проблема: автоматическое распознавание регуляторных областей



Актуальность: знание структуры = контроль работы гена

Существующие методы распознавания сайтов: Weight Matrix, SITECON, Motif Finders, etc...



Недостаток: не учитывают взаиморасположение

Цель: разработка системы для иерархического анализа регуляторных областей



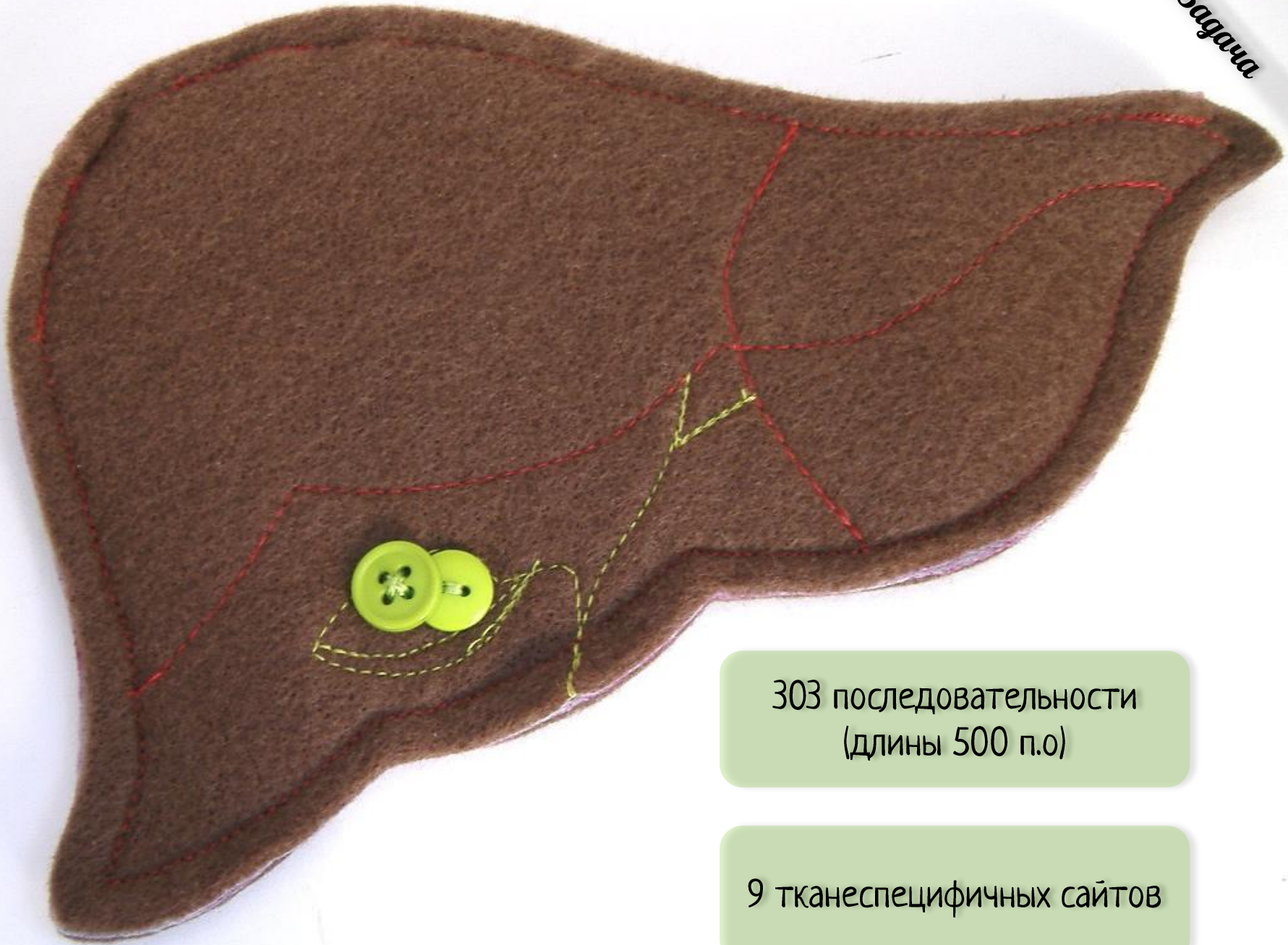
# Структура и требования

*Автоматическое  
построение модели*

*Распознавание*

*Визуализация и  
интерфейс*

*Биологические  
результаты*



303 последовательности  
(длины 500 п.о)

9 тканеспецифичных сайтов



# UGENE

## Интегрированные инструменты молекулярного биолога

UGENE - [Workflow Designer - New schema]

File Actions Settings Tools Window Help

100% Element style Run mode Scripting mode

Elements Samples

Name filter:

Transcription Factor

- Build Frequency Matrix
- Build SITECON Model
- Build Weight Matrix
- Convert Frequency Matrix
- Read Frequency Matrix
- Read SITECON Model
- Read Weight Matrix
- Search for TFBS with SITECON
- Search for TFBS with Weight Matrix
- Write Frequency Matrix
- Write SITECON Model
- Write Weight Matrix

Utils

Custom Elements with Script

2: Tasks 3: Log

```
graph LR; A[Read sequence] -- Sequence --> C[Search for TFBS with SITECON]; B[Read SITECON model] -- Sitecon model --> C; C -- SITECON annotations --> D[Write Genbank];
```

**Read sequence**  
Reads sequence(s) from [negative\\_50.fa](#).

**Read SITECON model**  
Read model(s) from [the list of files](#).

**Search for TFBS with SITECON**  
For each sequence from [Read sequence](#), search transcription factor binding sites (TFBS) with all profiles provided by [Read SITECON model](#). Recognize sites with [similarity 80%](#), process [both strands](#). Output the list of found regions annotated as [misc feature](#).

**Write Genbank**  
Save all sequences from [Read sequence](#) to [negative\\_mrk.gb](#).

Property Editor

Element name: Write Genbank

**Write Sequence** : Writes all supplied sequences to file(s) in selected format.

To configure the parameters of the element go to "Parameters" area below.

Parameters

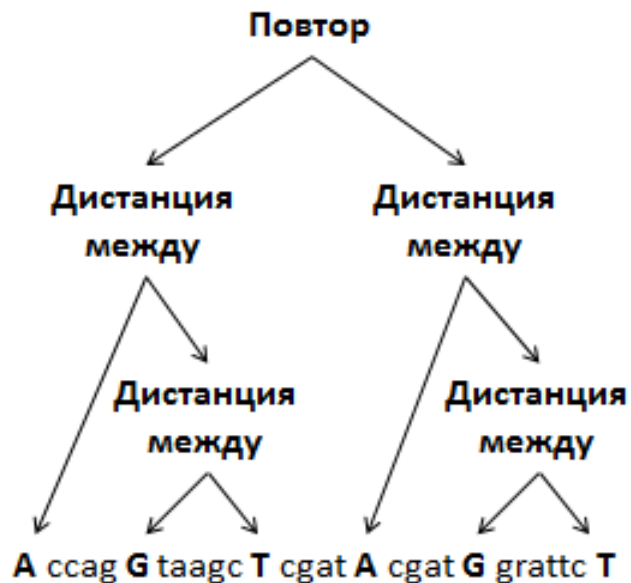
Name	Value
Accumulate objects	True
Document format	genbank
Split sequence	1
Output file	D:/dev/QT/Test ...negative_mrk.gb
Existing file	Rename

Input data

Set of annotations	<List of values>
Sequence	Sequence (by Read sequence)
Location	<empty>

No active tasks

# Комплексный сигнал



Алгоритм: Семантический вероятностный вывод

U - UGENE - [Expert Discovery]

File Actions Settings Tools Window Help

Items

- Sequences
  - Positive
  - Negative
  - Control
- Markup
  - MOTIFS
  - UGENE ANNOTATION
- Complex signals
  - 11
  - 22
    - NewSignal152
    - NewSignal150
    - NewSignal78
    - NewSignal52
    - NewSignal80
    - NewSignal0
    - NewSignal151
    - NewSignal149
    - NewSignal1
    - NewSignal6**
      - \$.s Distance from 0 to 10 taking into account order
        - "AAAATA" from family "MOTIFS"
        - "AAAAAT" from family "MOTIFS"
      - NewSignal85

Editor

Name	NewSignal6
Description	
<b>General information</b>	
Probability	48.7179% (19 / 39)
Pos. coverage	6.99588% (17 / 243)
Neg. coverage	0.781893% (19 / 2430)
Fisher	1.24281e-10
ul	0.715335

NM\_000014 | 5,11969\_neg7 [dna]

empty

1 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420 440 460 480 500

GGGCAACGTATAGCGTTGAACAAGCGACATCATCACAAAGTAGAAGCTACAGATCCCTTTGCAGCTTTCATTACA

12 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74

NM\_000014 | 5,11969\_neg0 [dna]

rec data (2)

1 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420 440 460 480 500

GCTGTTTACTTAGTCTAAGCTCTTGAATGCTGTGTTTGATATTTGGGTCTCATAGCCACGTATACATACCTTTT

12 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74

NM\_000128 | 12,6442 [dna]

Score graph [1, 500], Window: 2, Step 1

5.399

AGCGTACCTGAAGGGAAACTGACTGGGCAAGAGCCAGGCTCTTAGGAGCTGTCTGTAAGCTTTCATCCATCCG

260 265 270 275 280 285 290 295 300 305 310 315 320 325 330 333

Name Value

- Auto-annotations [Negative | NM\_000014 | 5,1...
- Auto-annotations [Negative | NM\_000014 | 5,1...

2: Tasks 3: Log No active tasks

UGENE - [Expert Discovery]

File Actions Settings Tools Window Help

Items

- Sequences
  - Positive
  - Negative
  - Control
- Markup
  - MOTIFS
  - UGENE ANNOTATION
- Complex signals
  - 11
  - 22
    - NewSignal152
    - NewSignal150
    - NewSignal78
    - NewSignal52
    - NewSignal80
    - NewSignal0
    - NewSignal151
    - NewSignal149
    - NewSignal1
    - NewSignal6
      - Distance from 0 to ...
      - "AAAATA" from ...
      - "AAAAAT" from ...
    - NewSignal85

Editor

Name	Description
NewSignal6	

General information

Probability	48.7179% (19)
Pos. coverage	6.99588% (17)
Neg. coverage	0.781893% (19)
Fisher	1.24281e-10
ul	0.715335

### Setup Recognition Bound

Recognition Bound: 21,600

Optimize Recognition Bound

Information

Probability of negative sequence recognition: 0.000411523

Probability of positive sequence rejection: 0.473251

Recognition Graph

Minimum Bound: 0 Maximum Bound: 40

Bound Step: 0,1

Recalculate Graph Values

OK Cancel

380 400 420 440 460 480 500

380 400 420 440 460 480 500

CCTTTGCAGCTTTCATTACA

56 58 60 62 64 66 68 70 72 74

380 400 420 440 460 480 500

380 400 420 440 460 480 500

GCCACGTATACATACCTTTT

56 58 60 62 64 66 68 70 72 74

380 400 420 440 460 480 500

5.399

rec.d

rec.data

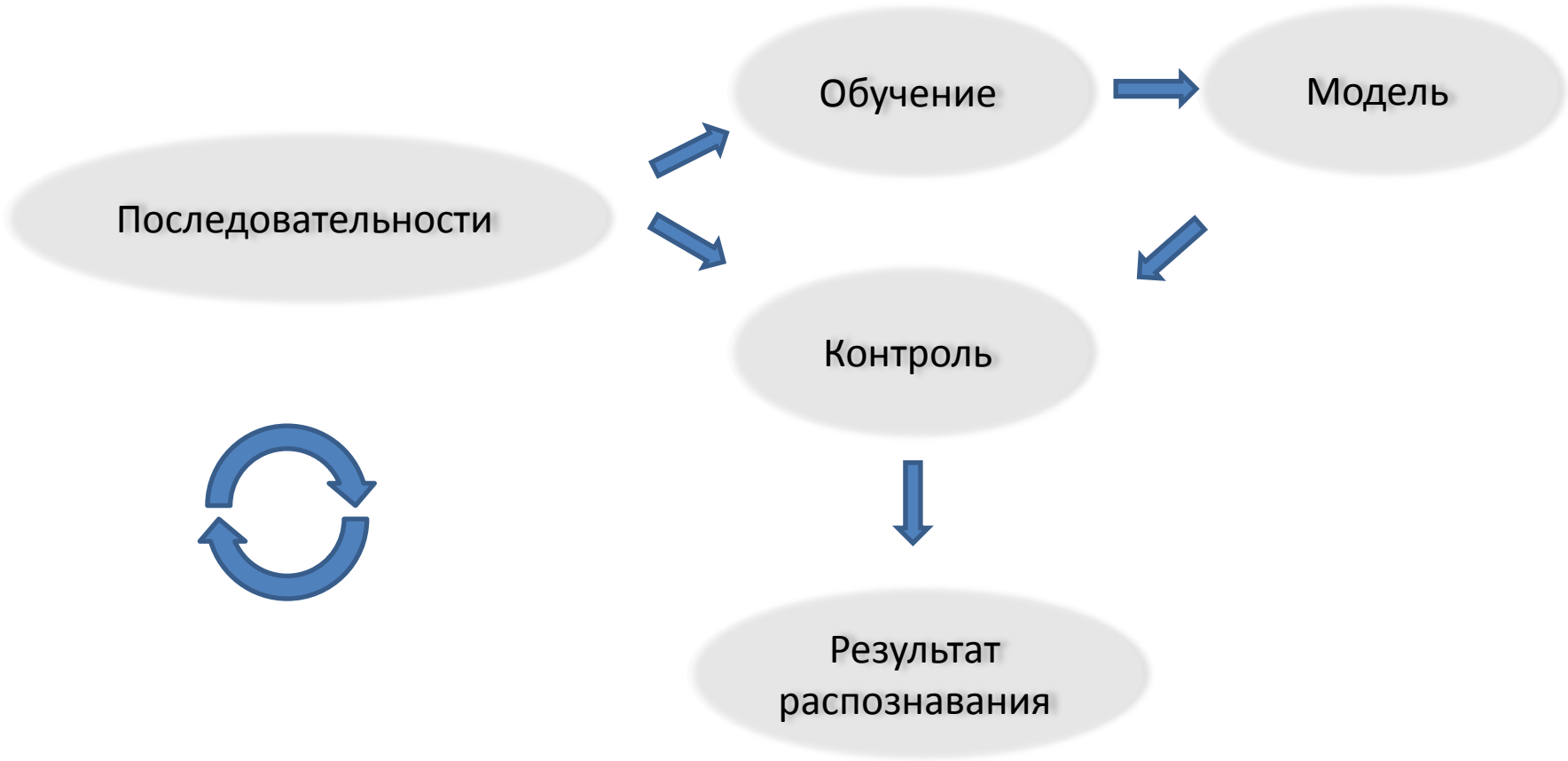
380 400 420 440 460 480 500

TGTACTTTGCTTCCCATCCG

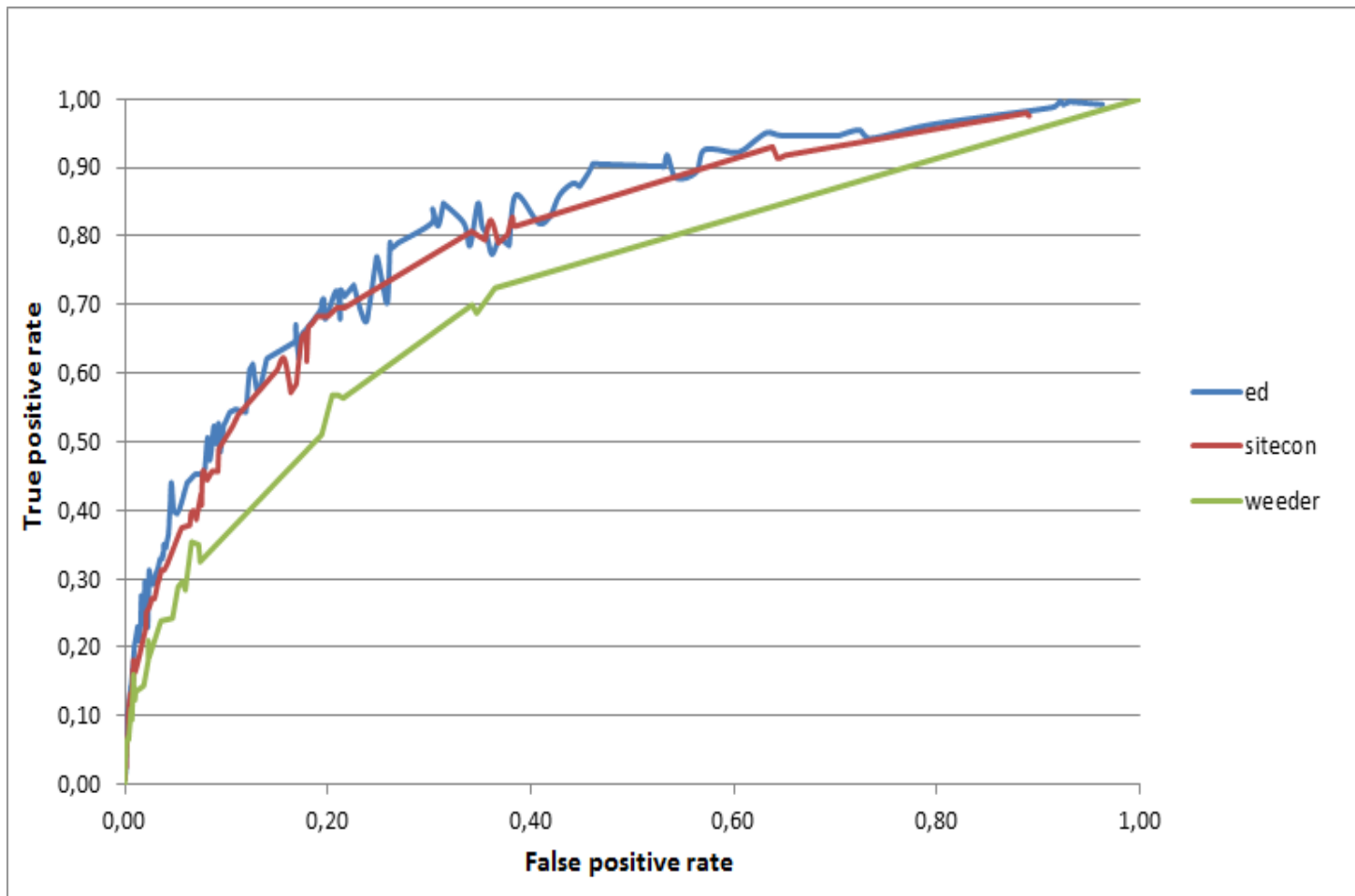
315 320 325 330 333

No active tasks

# Проверка модели: Cross-validation



# Receiver operating characteristic



# Результаты

Система, способная:

- Интегрировать сигналы разных типов
- Качественно распознавать
- Генерировать значимые результаты
- Отображать данные различных типов



## Публикации

1. “Анализ последовательностей регуляторных районов генов реляционной системой ExpertDiscovery, встроенной в пакет UGENE”. “Вестник НГУ”, Том 10, Выпуск 1, Новосибирск, 2011
2. “ExpertDiscovery and UGENE integrated system for intelligente analysis of regulatory regions of genes”. In *Silico Biol.* 2011-2012;11(3-4):97-108. doi: 10.3233/ISB-2012-0448.
3. “Analysis and Prediction of Regulatory Regions of Eukaryotic Genes by integrated UGENE and ExpertDiscovery Systems”. Special issue: ECML/PKDD 2011 (5th Workshop on Data Mining in Functional Genomics and Proteomics: Current Trends and Future Directions), Athens, 2011

## Конференции

1. I Международная научная студенческая конференция «Студент и научно-технический прогресс» (Диплом III степени);
2. Международная конференция «Современные проблемы математики, информатики, биоинформатики»;
3. Международная конференция «Постгеномные методы в биологии, лабораторной и клинической медицине: геномика, протеомика, биоинформатика».
4. VIII Международная конференция по биоинформатике, системной биологии, регуляции и структуре геномов.





