РАЗРАБОТКА ТЕХНОЛОГИИ РЕКОНСТРУКЦИИ И КОМПЬЮТЕРНОГО АНАЛИЗА ГЕННЫХ СЕТЕЙ И ЕЕ ПРИМЕНЕНИЕ В БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Соискатель: *Ананько Елена Анатольевна*

Научный руководитель:
член-корр. РАН, д.б.н., профессор
Колчанов Николай Александрович

Рецензенты:
д.б.н. *Меркулова Татьяна Ивановна* (ИЦиГ)
к.б.н. *Савинкова Людмила Кузьминична* (ИЦиГ)

Актуальность проблемы

Бурное развитие экспериментальных технологий в области молекулярной биологии и генетики привело к появлению огромных объемов информации. При этом ни один экспериментальный метод, независимо от его эффективности, сам по себе не может дать комплексного представления о биологическом объекте.

Острая потребность в осмыслении больших массивов информации привела к появлению нового научного направления - системной компьютерной биологии.

Центральным понятием и основным объектом изучения системной компьютерной биологии являются *генные сети* - молекулярно-генетические системы, обеспечивающие формирование разнообразия фенотипических характеристик организмов на основе информации, закодированной в их геномах.

Актуальность проблемы

Классический пример сложно организованной генной сети представляет собой интерфероновая система. Интерфероны - это основные регуляторы иммунного, противовирусного и противобактериального ответов.

Исследование генной сети интерфероновой системы с помощью современных компьютерных технологий может открыть пути к созданию новых лекарственных препаратов с более точно направленным воздействием и минимумом побочных эффектов, а также стимуляторов иммунной системы и других биологически активных веществ.

Эти исследования имеют важное практическое значение, поскольку раскрывают причины возможных нарушений регуляции целого ряда жизненно важных функций организма и позволяют приблизиться к решению проблемы их генетической коррекции.

Цель работы:

разработка технологии компьютерной реконструкции и анализа генных сетей;

исследование генных сетей интерфероновой индукции противовирусного ответа

и построение методов распознавания интерферон-индуцируемых генов эукариот

Поставленные задачи:

1. Создание технологии формализованного описания генных сетей, включающей в себя ***принципы описания*** отдельных классов объектов сетей, ***методы*** реконструкции генной сети *in silico*, ***словари*** терминов, ***базу данных*** и ***программные средства*** для поддержки базы, визуализации и анализа данных
2. Анализ особенностей организации генных сетей эукариот на основе информации, накопленной в общей базе данных по генным сетям
3. Создание базы данных по генным сетям интерфероновой индукции противовирусного ответа у эукариот
4. Компьютерный анализ особенностей организации и функционирования генных сетей интерфероновой индукции у млекопитающих

Поставленные задачи:

5. Построение методов распознавания сайтов связывания транскрипционных факторов, важных для функционирования генной сети интерфероновой системы, а именно, ISGF3, IRF1, STAT1, NF-κB, AP1

6. Определение характерных для интерферон-индуцируемых генов закономерностей в расположении сайтов связывания разных транскрипционных факторов

7. Разработка методов распознавания интерферон-индуцируемых промоторов и энхансеров в геномах эукариот

8. Поиск потенциальных интерферон-индуцируемых генов человека

Задача 1

Создание технологии формализованного описания генных сетей, включающей в себя

принципы описания отдельных классов объектов сетей,

методы реконструкции генной сети *in silico*,

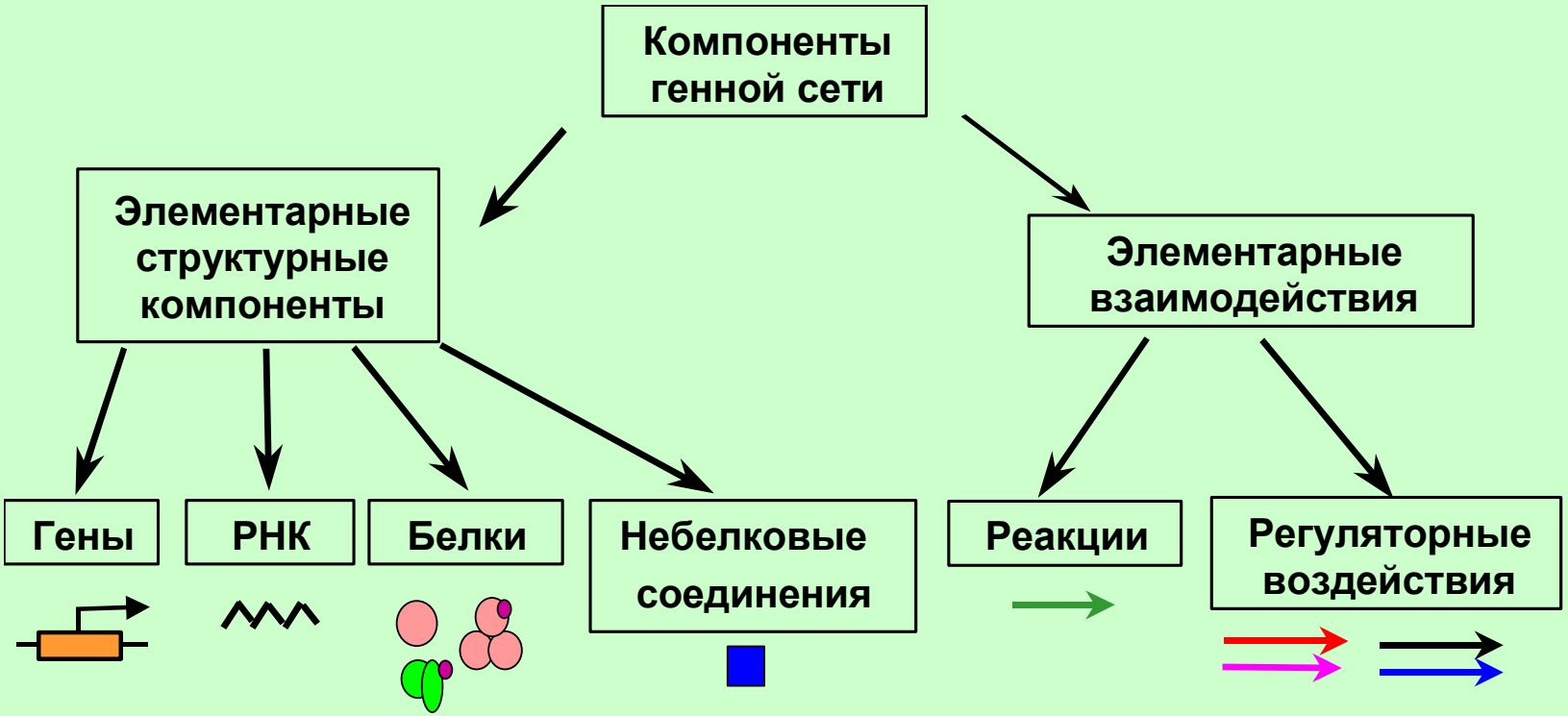
словари терминов,

базу данных и

программные средства для поддержки базы, визуализации и анализа данных

Задача 1

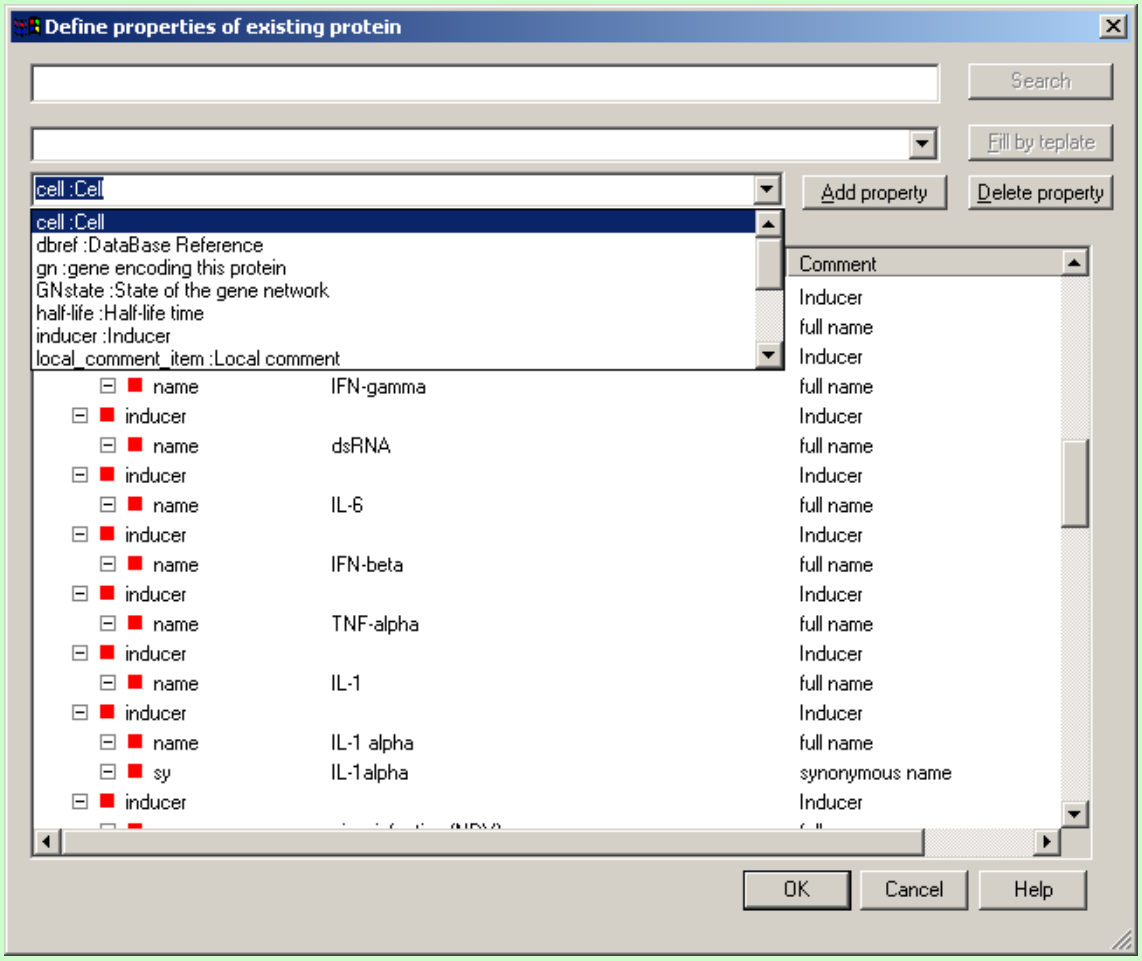
принципы описания отдельных классов объектов сетей



Задача 1

методы реконструкции генной сети *in silico*

1. **Информация об элементарных объектах и взаимодействиях между ними берется из опубликованных научных статей и распределяется по информационным полям соответствующих таблиц базы данных**



<reaction id="55">
<object id="reaction1ADE5B68-DE2C-48C6-B30D-4E56D137E586">

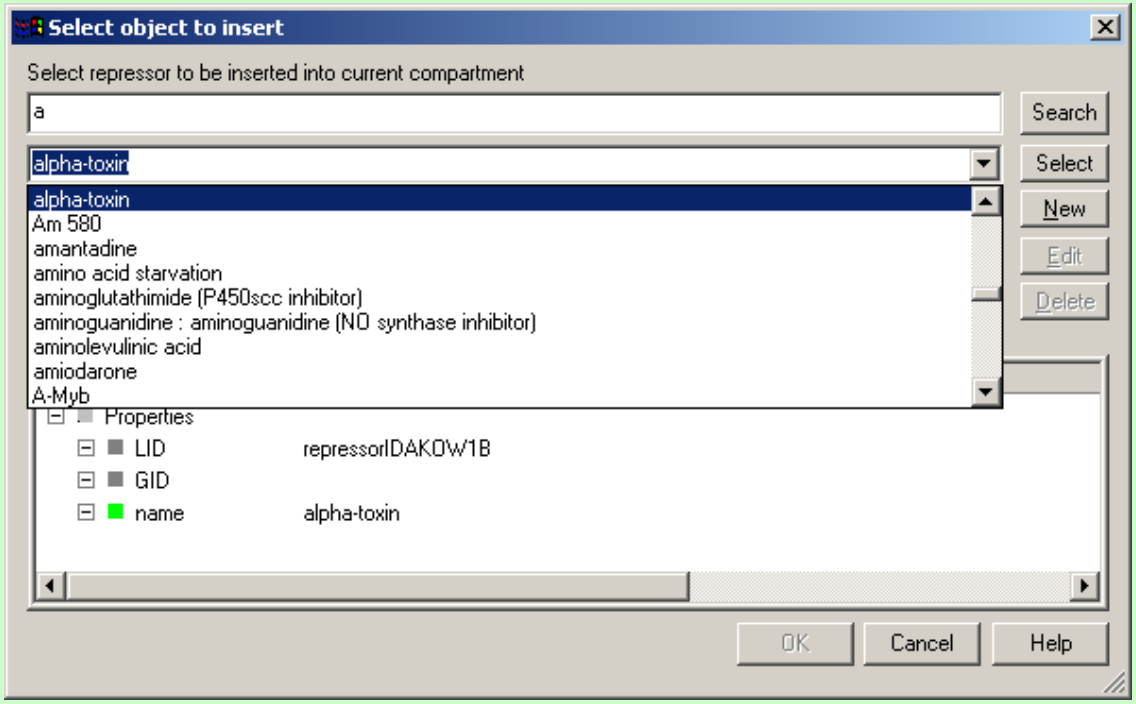
<type>irreversible</type>

<nm>reaction</nm><ef>indirect</ef><input><input_item><gene id="geneID17155"><name>interferon regulatory factor 2</name><os id="organismID953"><Latin>Homo sapiens</Latin><English>human</English></os></gene></input_item></input><output><output_item><protein id="proteinID32708"><name>interferon regulatory factor-2</name><os id="organismID953"><Latin>Homo sapiens</Latin><English>human</English></os></protein></output_item></output><ref id="literID23092"><authors>Harada H., Takahashi E.-I., Itoh S., Harada K., Hori T.-A. and Taniguchi T.</authors><journal id="journal234"><nm>Mol.Cell.Biol.</nm></journal><title>Structure and regulation of the human interferon regulatory factor 1 (IRF-1) and IRF-2 genes: implications for a gene network in the interferon system.</title><year>1994</year><pages><first>1500</first><last>1509</last></pages></ref><dt>30.3.1999.;Ananko E.;created</dt><dt>19.02.2005;Ananko E.A.;edited</dt></object>
<x>434.647839</x>
<y>1001.386898</y>
<expanded>true</expanded>
</reaction>

Задача 1

методы реконструкции генной сети *in silico*

2. **Максимальная верификация данных с помощью различных словарей**



В систему встроено более 20 словарей

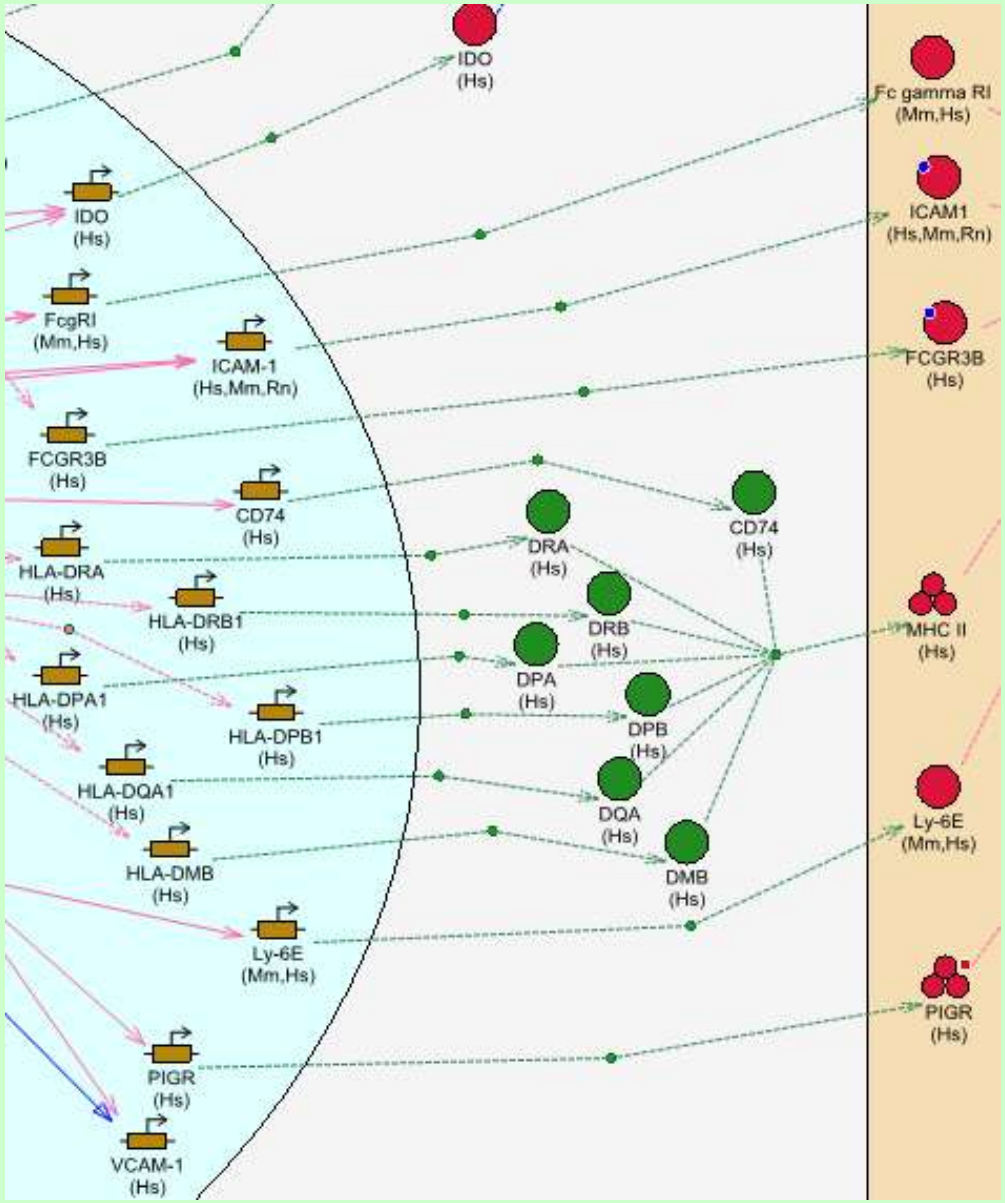
содержащих в общей сложности более 1000 терминов

Примеры словарей:

- **тканей и органов**
- **типов клеток и клеточных линий**
- **индукторов и репрессоров**
- **названий журналов**
- **названий баз данных**

Задача 1

методы реконструкции генной сети *in silico*



3. **Каждый элементарный объект и взаимодействие имеют пространственную привязку к клеточному компартменту**

Задача 1

методы реконструкции генной сети *in silico*

4. Иерархия уровней представления генной сети

1. **Три уровня представления генной сети:**

Молекулярный

На этом уровне описываются взаимодействия молекул в пределах, как правило, одного-двух компартментов клетки. Например, пути передачи сигналов или регуляция транскрипции. Детализация процессов на этом уровне максимальна.

2. ***Клеточный***

На этом уровне описываются процессы, протекающих в различных компартментах клетки, и их влияние друг на друга. Детализация описания меньше, чем на первом уровне. Многие реакции и регуляторные события здесь описываются как непрямые.

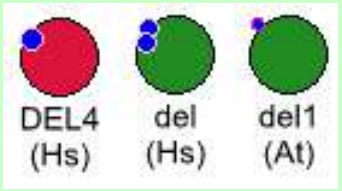
Организменный

На этом уровне описываются взаимодействия клеток, тканей, органов. Здесь же возможно представление взаимодействий разных организмов, например патологического микроорганизма и организма-хозяина, или симбиотических организмов. Детализация описания минимальна.

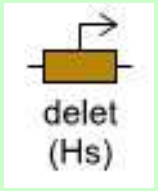
Задача 1

методы реконструкции генной сети *in silico*

Визуализация данных в виде двумерного графа:
отображение элементарных объектов

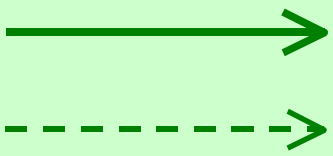


Белки

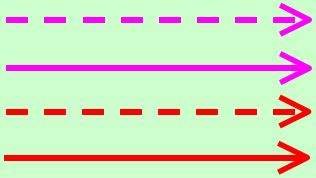


Гены, РНК и низкомолекулярные соединения

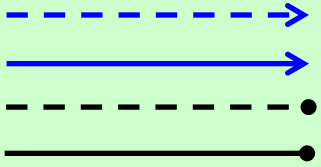
Взаимодействия между объектами:



Реакции



Положительные регуляторные воздействия

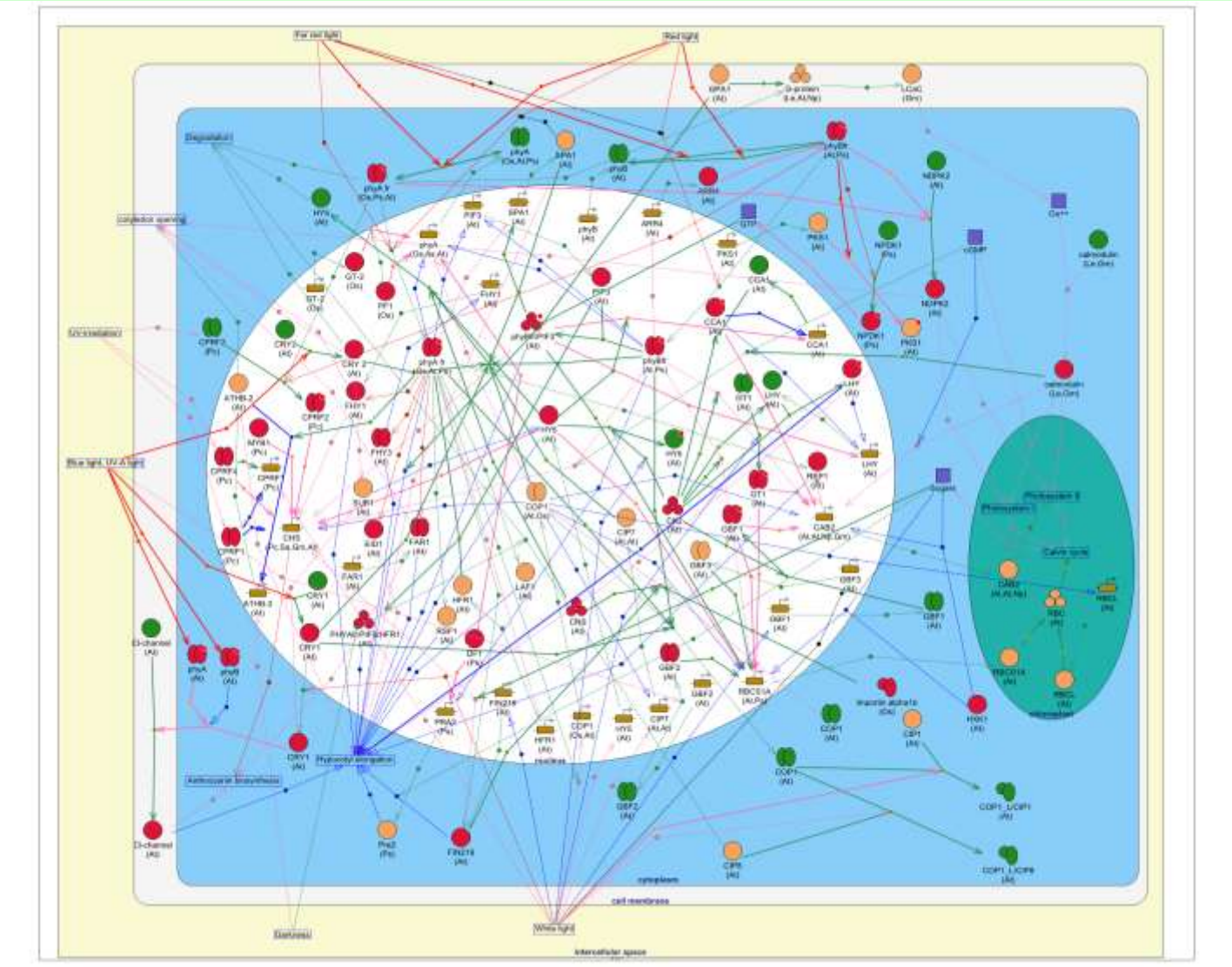


Отрицательные регуляторные воздействия

Задача 1

методы реконструкции генной сети *in silico*

Визуализация данных в виде двумерного графа



Задача 1

словари терминов и ***база данных***

База данных GeneNet уже несколько лет активно используется сотрудниками института при выполнении различных проектов. На ***1 ноября 2007 года*** в базе имелось описание:

42 генных сетей эукариот и **23** генных сетей прокариот

3711 белков

2112 генов

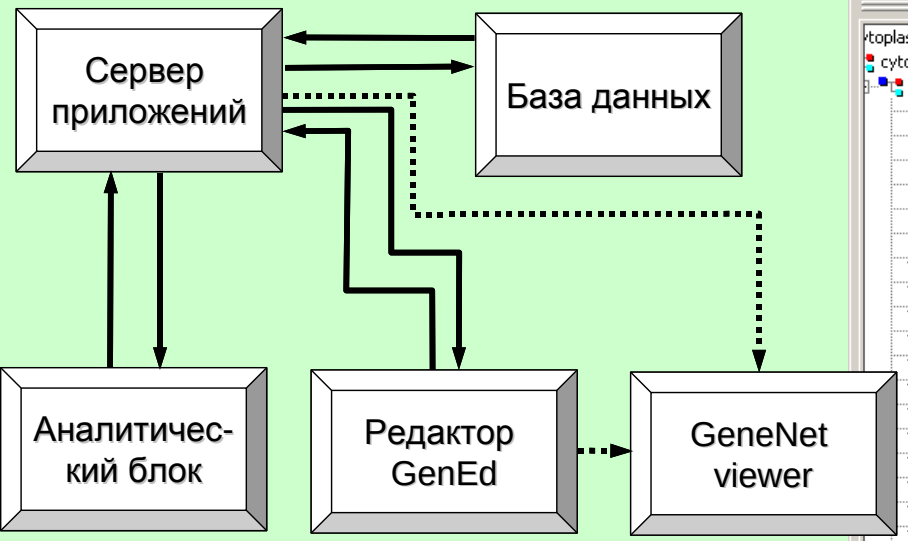
254 оперонов

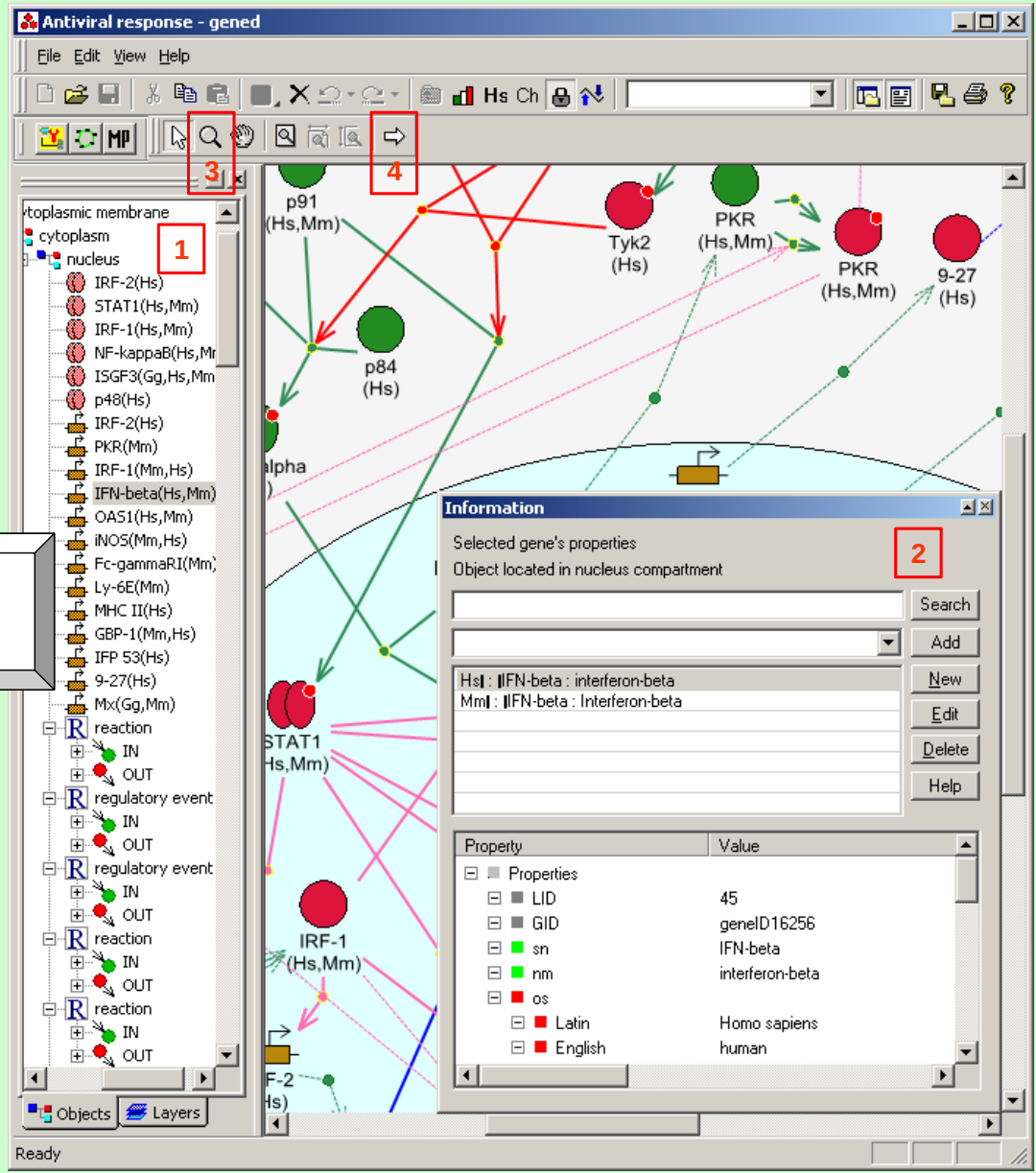
11 745 взаимодействий

аннотировано **8 755** научных публикаций

Задача 1

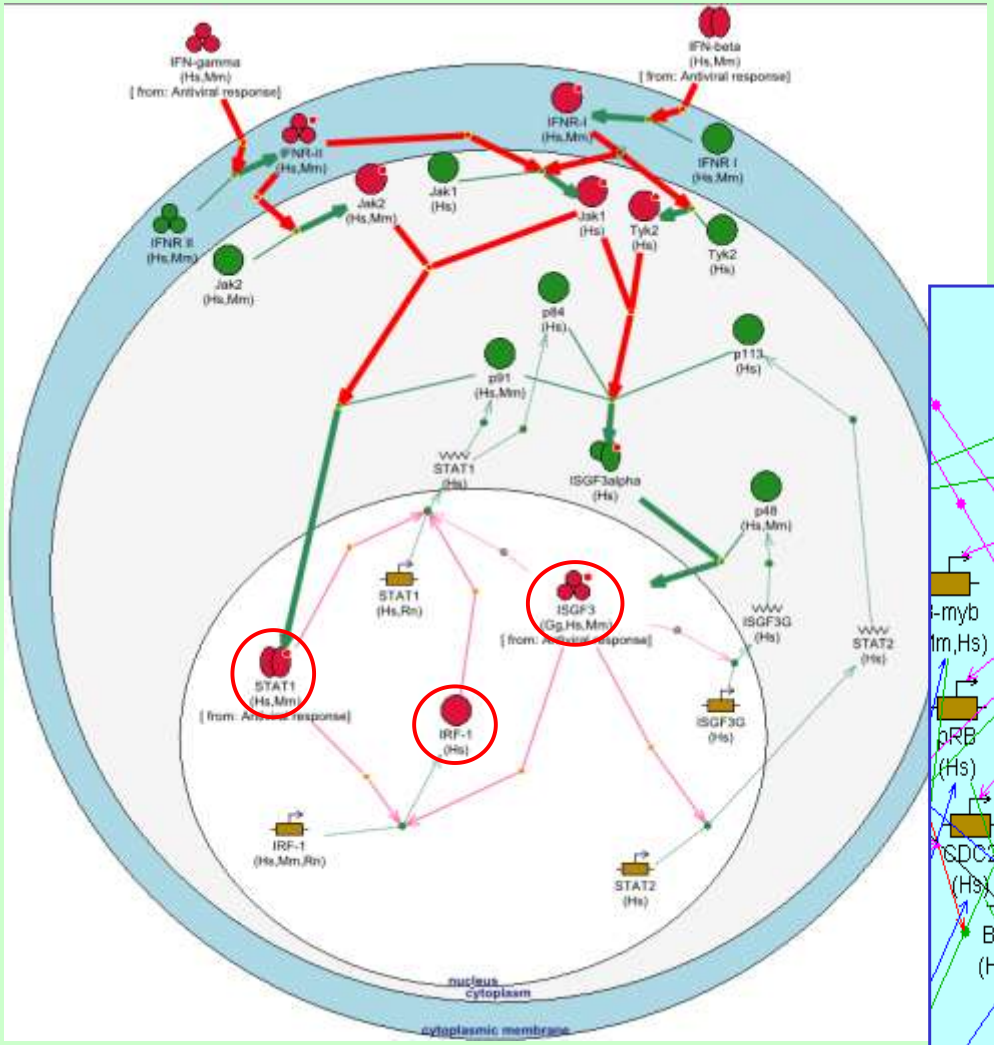
программные средства



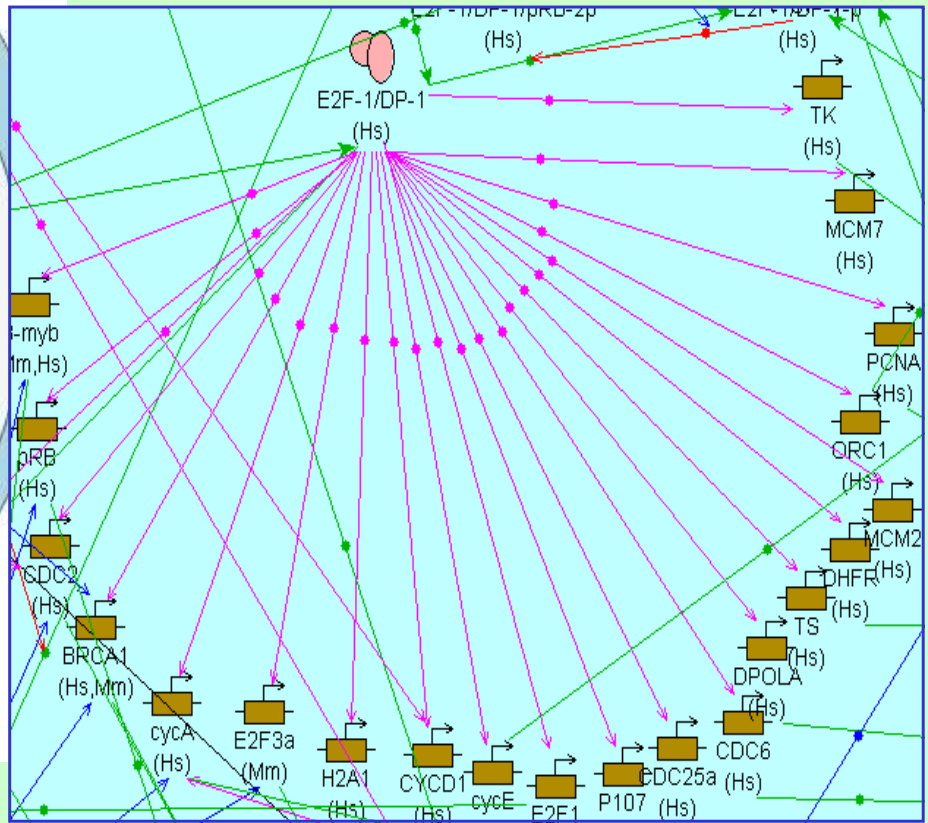


Задача 2

Анализ особенностей организации генных сетей эукариот на основе информации, накопленной в общей базе данных по генным сетям

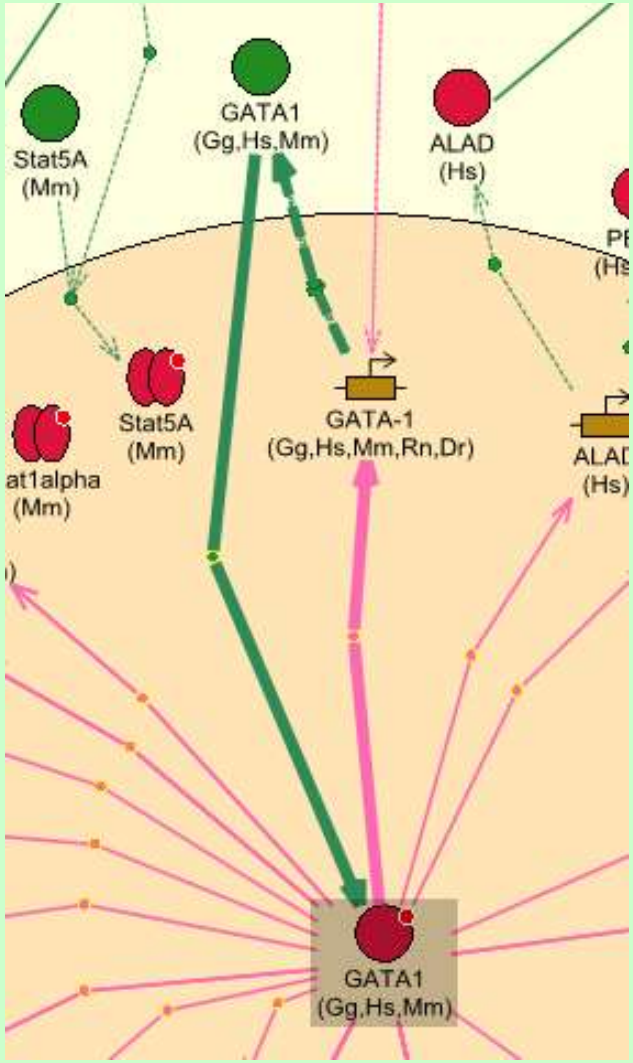


Пути передачи сигналов и ключевые регуляторы

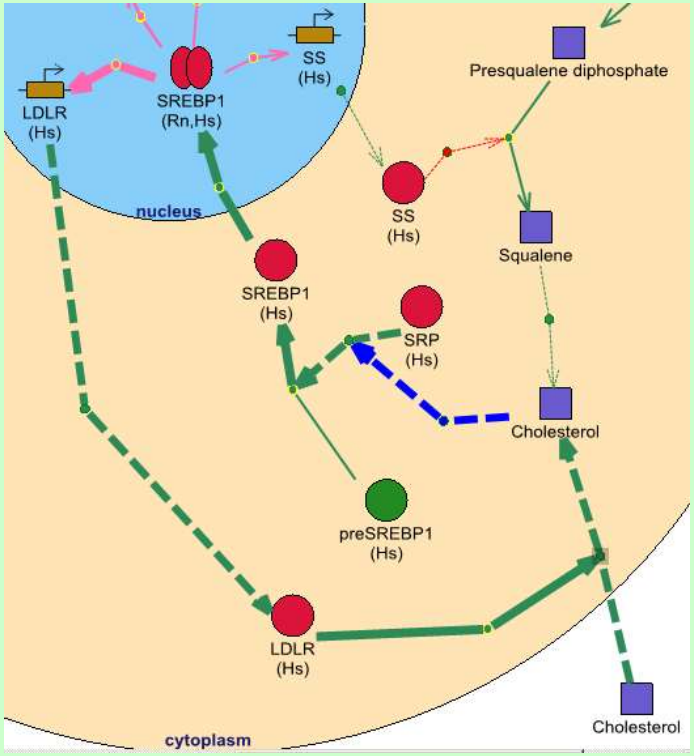


Задача 2

Анализ особенностей организации генных сетей эукариот на основе информации, накопленной в общей базе данных по генным сетям

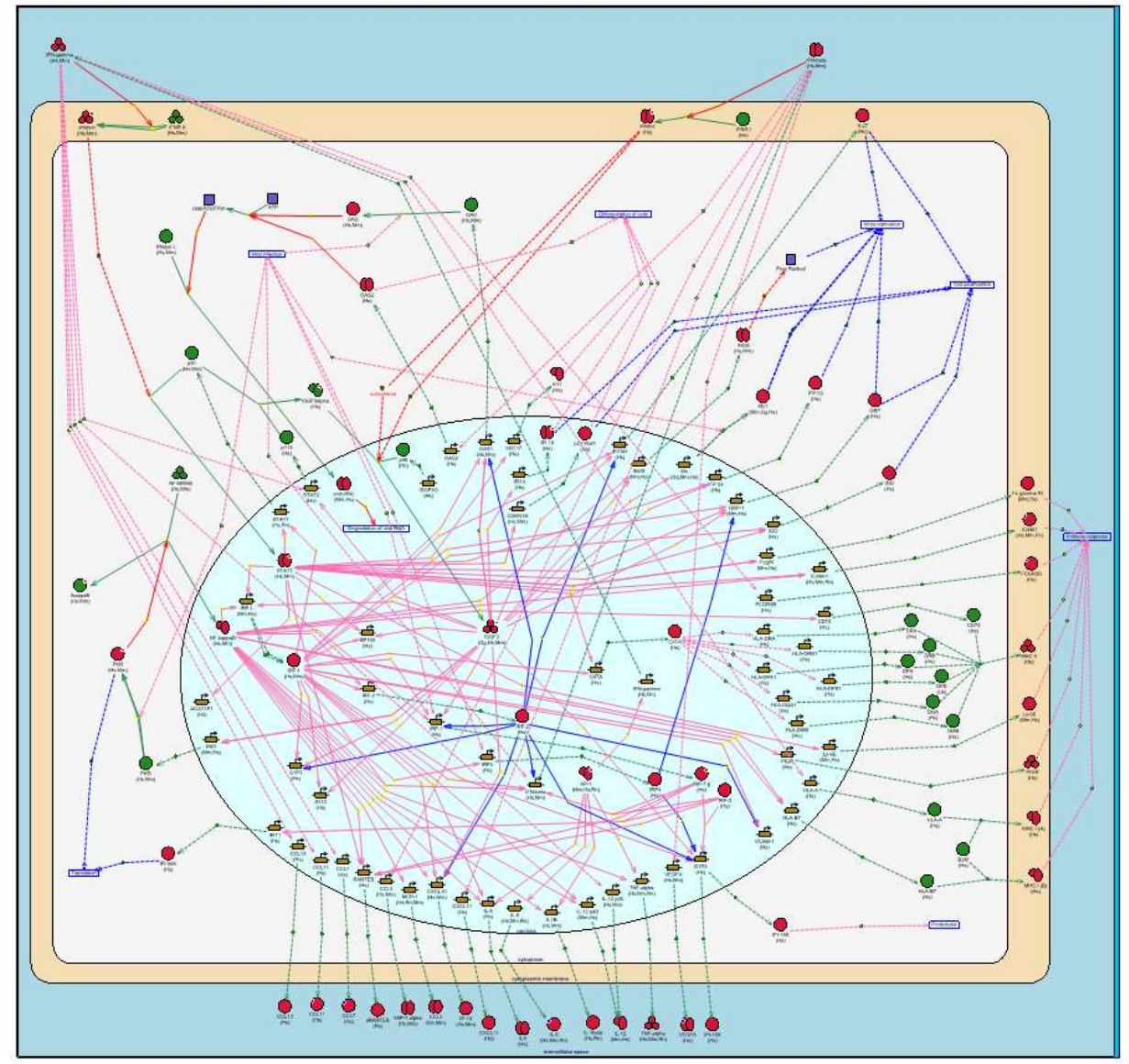


Регуляторные контуры с положительными или отрицательными обратными связями



Задача 3

Создание базы данных по генным сетям интерфероновой индукции противовирусного ответа у эукариот



Интерфероновая регуляция противовирусного ответа (генная сеть "Antiviral response")

108 белков

85 генов

219 взаимодействий

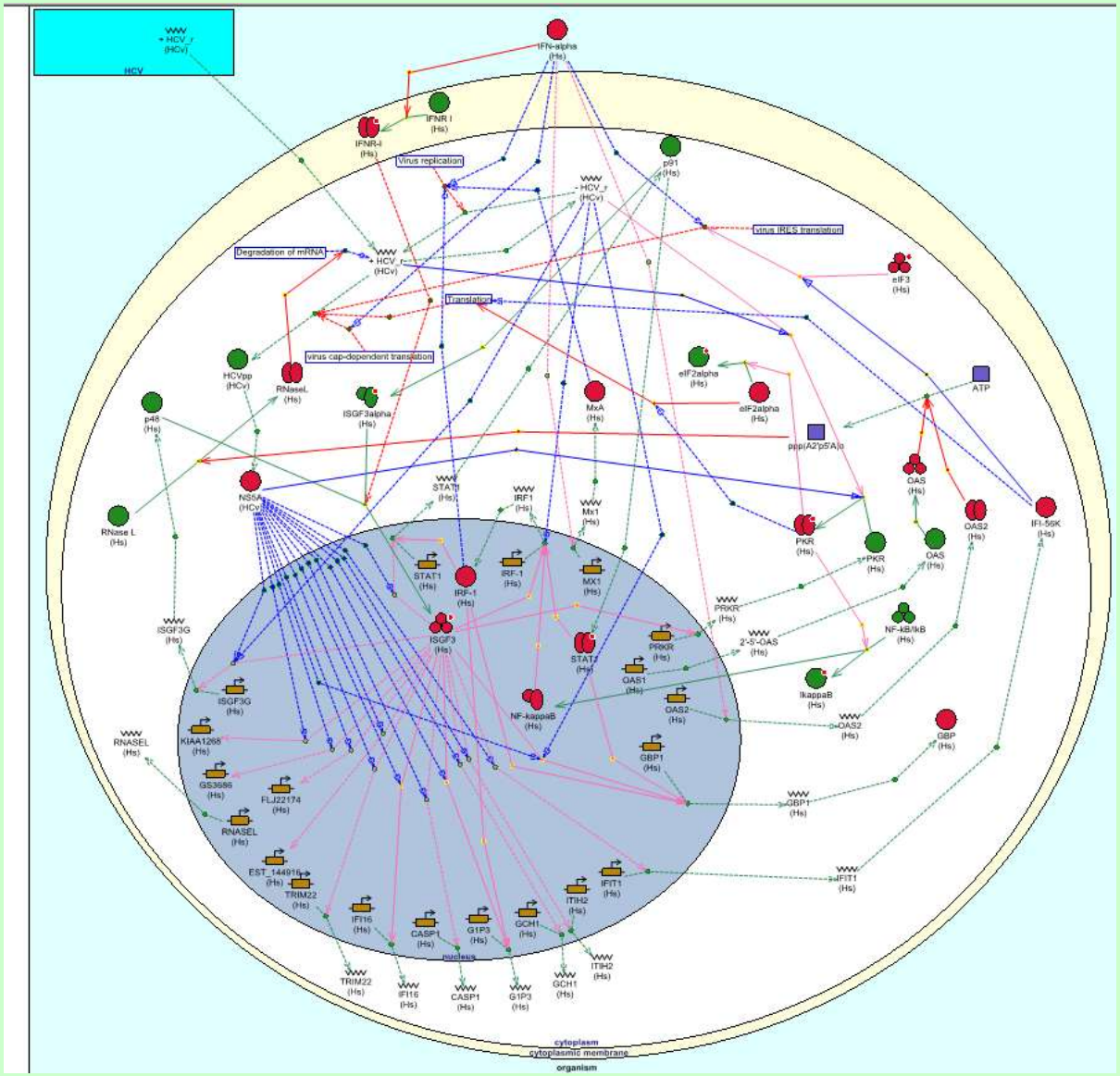
аннотировано 339 публикаций

Данные по 4 организмам:

человек, мышь, крыса, курица

Задача 3

Создание базы данных по генным сетям интерфероновой индукции противовирусного ответа у эукариот



Индукция противовирусного ответа интерфероном-α при гепатите С (генная сеть "Hepatitis C (IFN)")

27 белков

20 генов

107 взаимодействий

аннотировано 107 публикаций

Данные по 2 организмам:

человек,
M. tuberculosis

Задача 4

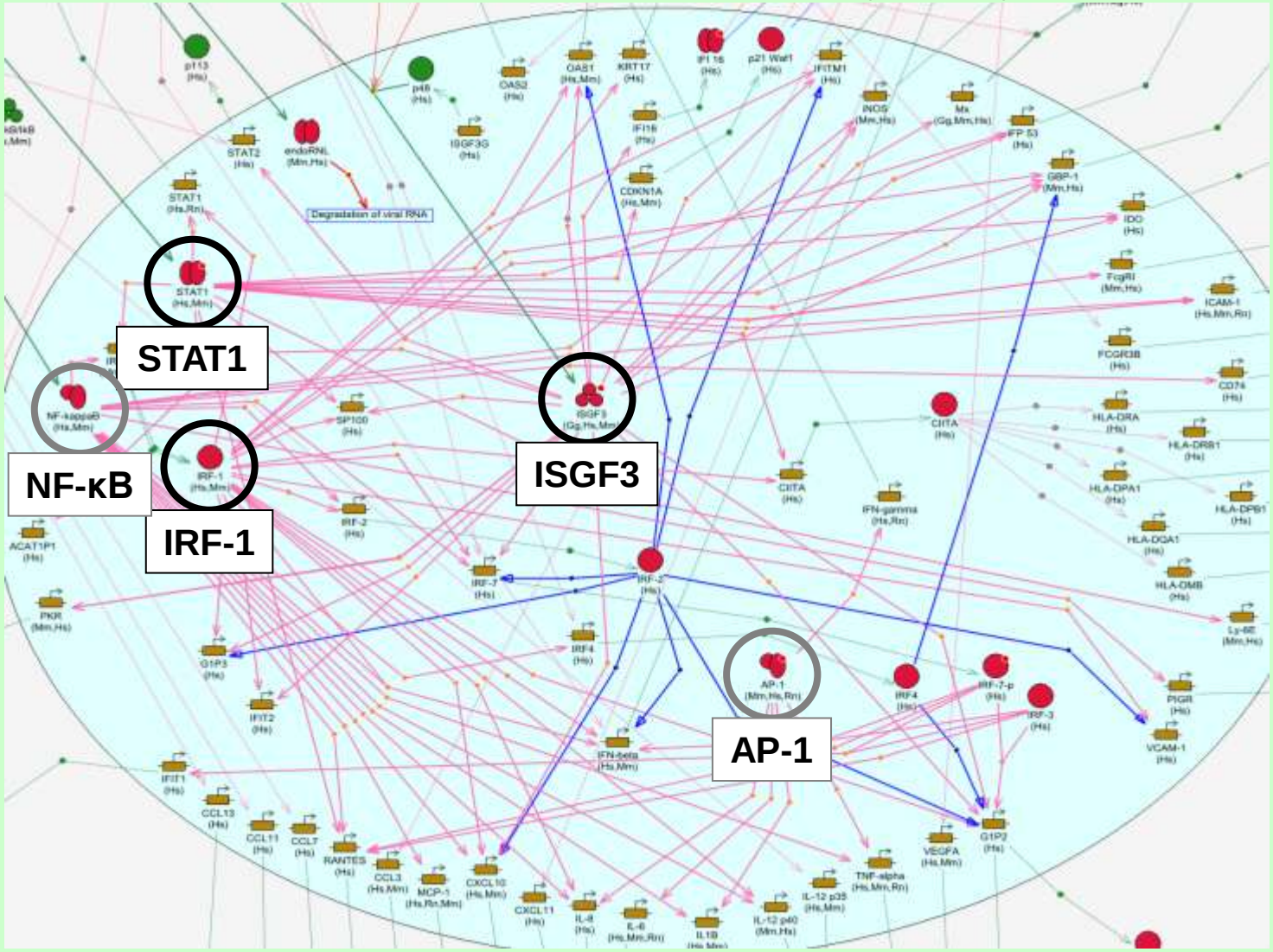
Анализ особенностей организации и функционирования генных сетей интерфероновой индукции у млекопитающих

Название схемы генной сети	Количество регуляторных циклов с положительной обратной связью	Количество регуляторных циклов с отрицательной обратной связью
Antiviral response	41	9
Hepatitis C (IFN)	37	5

Преобладание регуляторных циклов с положительной обратной связью

Задача 4

Анализ особенностей организации и функционирования генных сетей интерфероновой индукции у млекопитающих



Ключевые регуляторы – транскрипционные факторы

STAT1

ISGF3

IRF-1

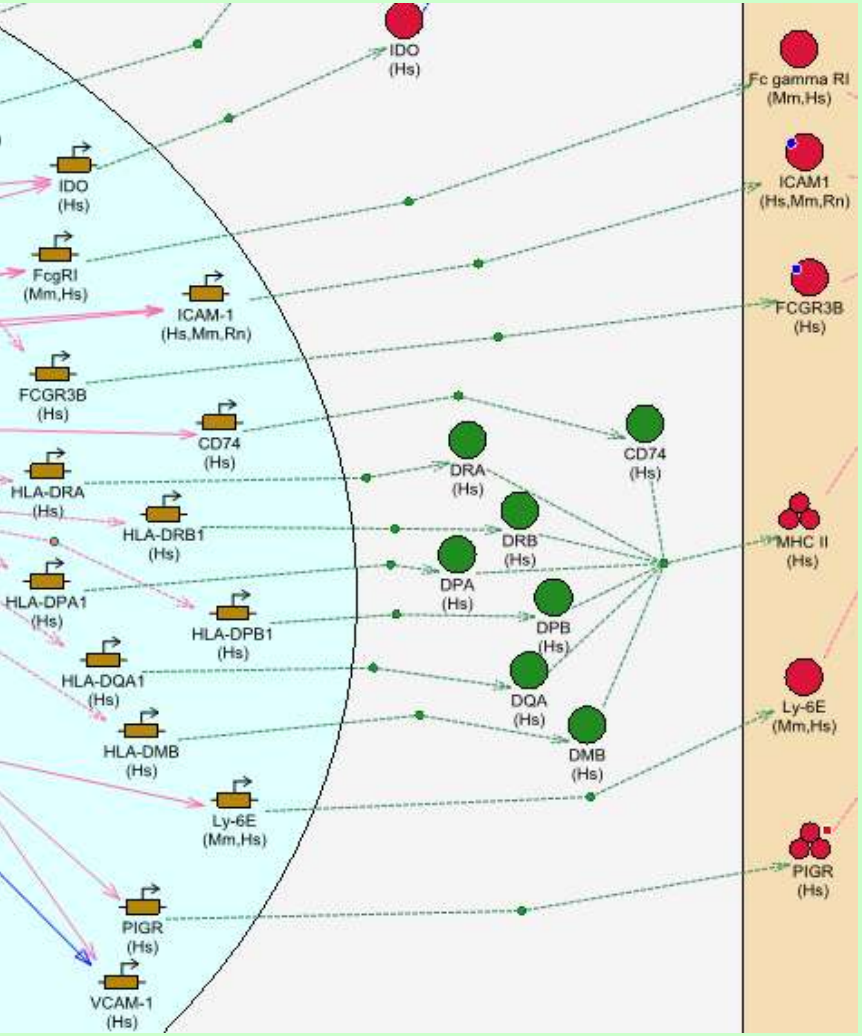
NF-κB

AP-1

Фрагмент генной сети "Antiviral response"

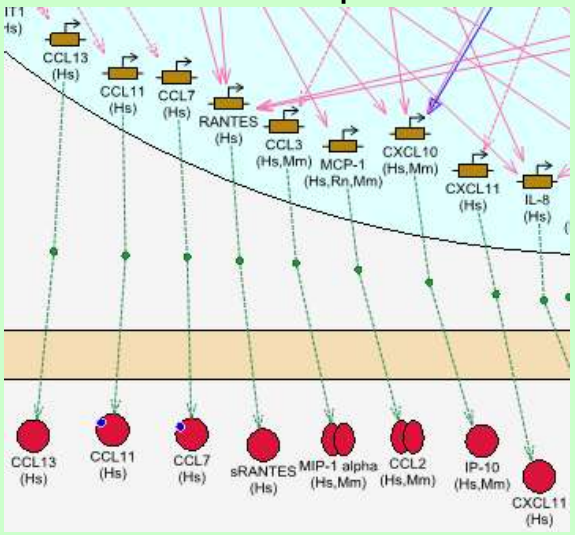
Задача 4

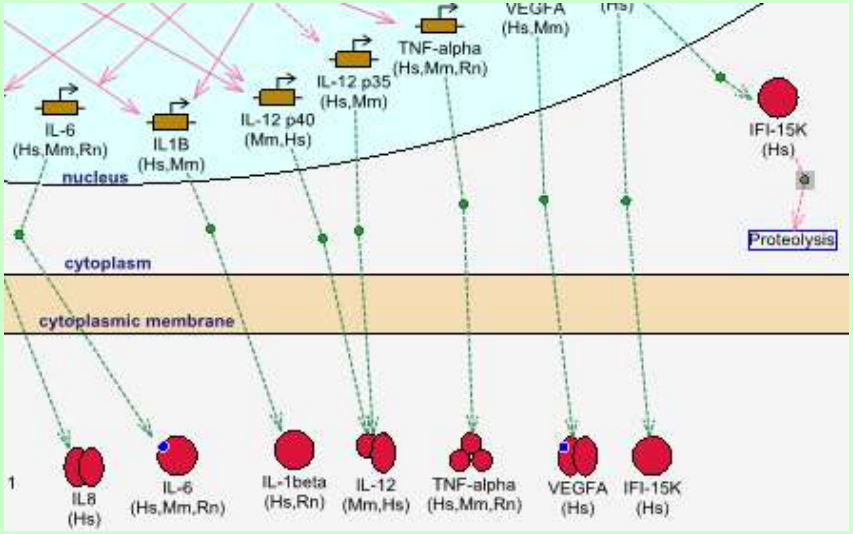
Анализ особенностей организации и функционирования генных сетей интерфероновой индукции у млекопитающих



Комплекс гистосовместимости, поверхностные рецепторы

Хемокины



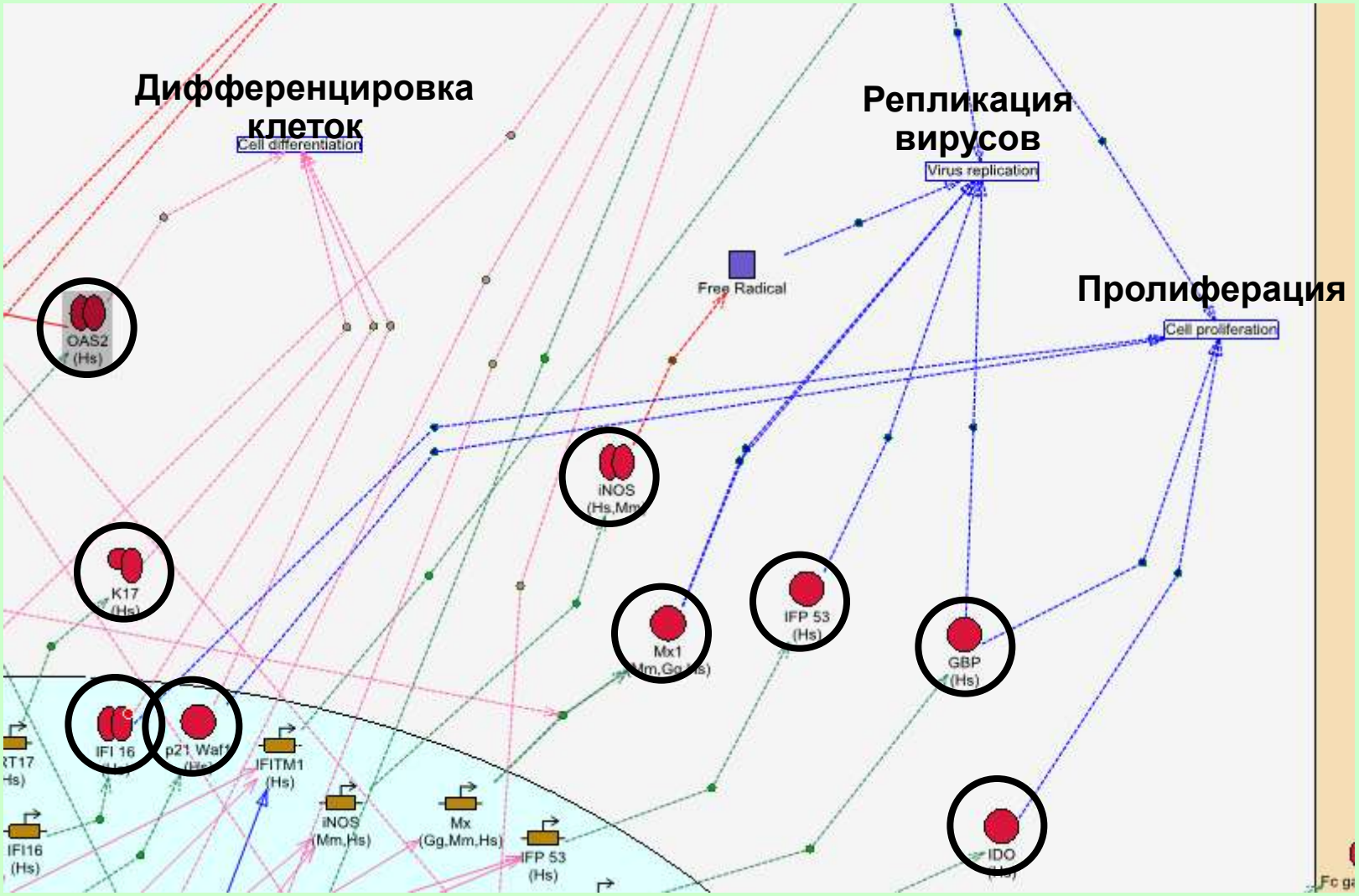


Цитокины

Фрагменты генной сети "Antiviral response"

Задача 4

Анализ особенностей организации и функционирования генных сетей интерфероновой индукции у млекопитающих



Фрагменты генной сети "Antiviral response"

Задача 5

Построение методов распознавания сайтов связывания транскрипционных факторов, важных для функционирования генной сети интерфероновой системы, а именно, ISGF3, IRF1, STAT1, NF-κB, AP1

Объем выборок, использованных для построения методов распознавания

Транскрипционный фактор	Кол-во сайтов
AP-1	20
NF-κB	44
ISGF3	24
STAT1	21
IRF-1	30

Длина последовательностей = 100 п.о., все сайты взяты из базы данных TRRD

Задача 5

Построение методов распознавания сайтов связывания ISGF3, IRF1, STAT1, NF-κB, AP1

Использовался три разных итерационных метода построения весовых матриц

Для каждой выборки было получено по 3 матрицы, которые, как правило, не совпадали между собой, но имели высокую степень сходства. Для распознавания отбиралась та матрица, которая обеспечивала наименьшую ошибку 2-го рода при фиксированной ошибке 1-го рода = 15%

Для снижения ошибки 2-го рода использовался метод статистического моделирования

Минимизация ложных предсказаний при условии, что ошибка 1-го рода не имела существенного увеличения

Задача 5

Построение методов распознавания сайтов связывания ISGF3, IRF1, STAT1, NF-κB, AP1

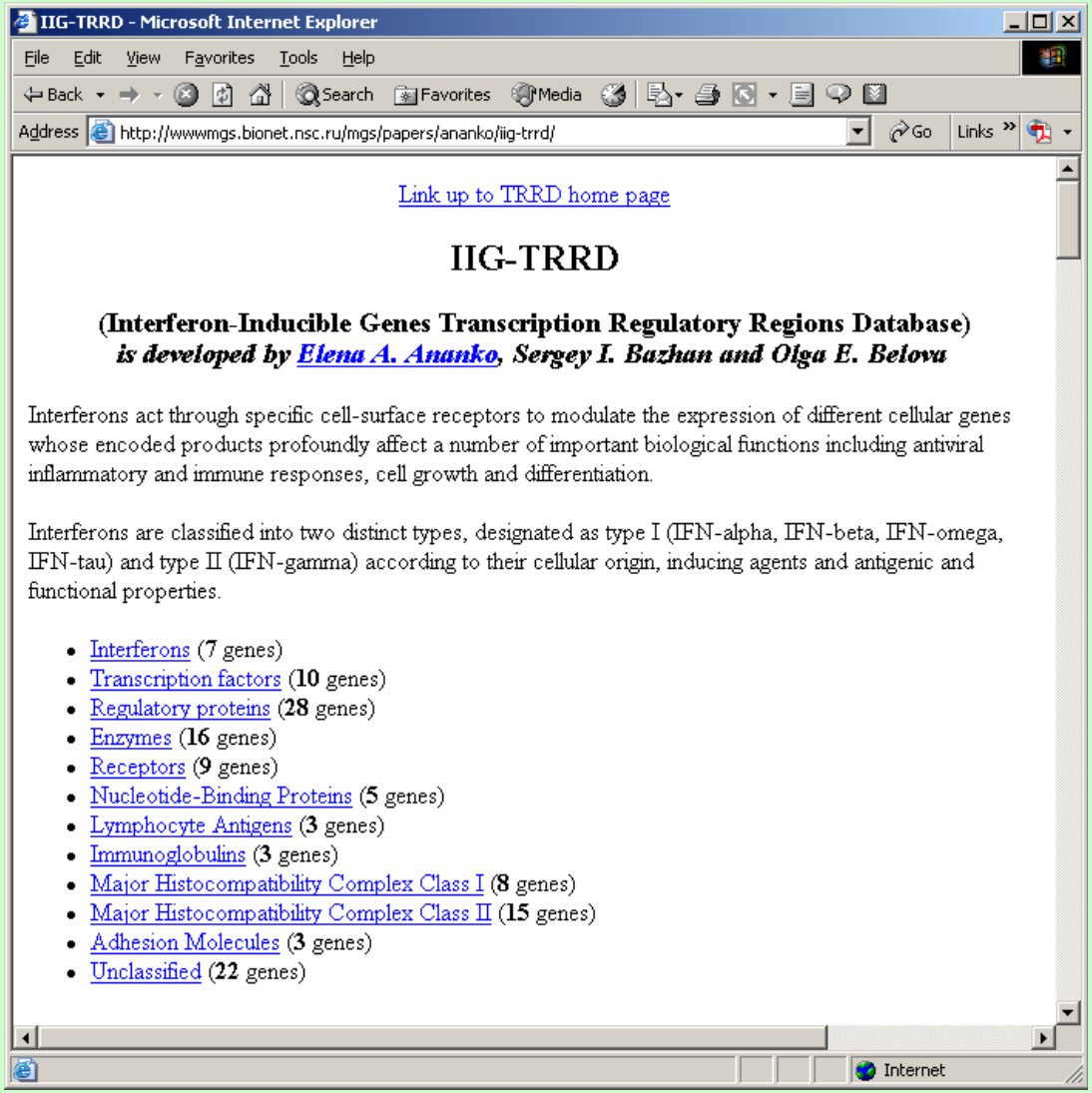
Характеристики полученных методов

ССТФ	Ошибка 1-го рода α1 (недопредсказание)	Ошибка 2-го рода α2 (перепредсказание)	Независимый контроль*
AP-1	37%	2.81E-04	нд*
IRF1	24%	9.59E-05	31.8%
ISGF3	25%	6.84E-04	46.2%
NF-κB	42%	5.32E-04	70.8%
STAT1	43%	8.82E-05	84.6%

* (уровень недопредсказания при заданной α_2)

Задача 6

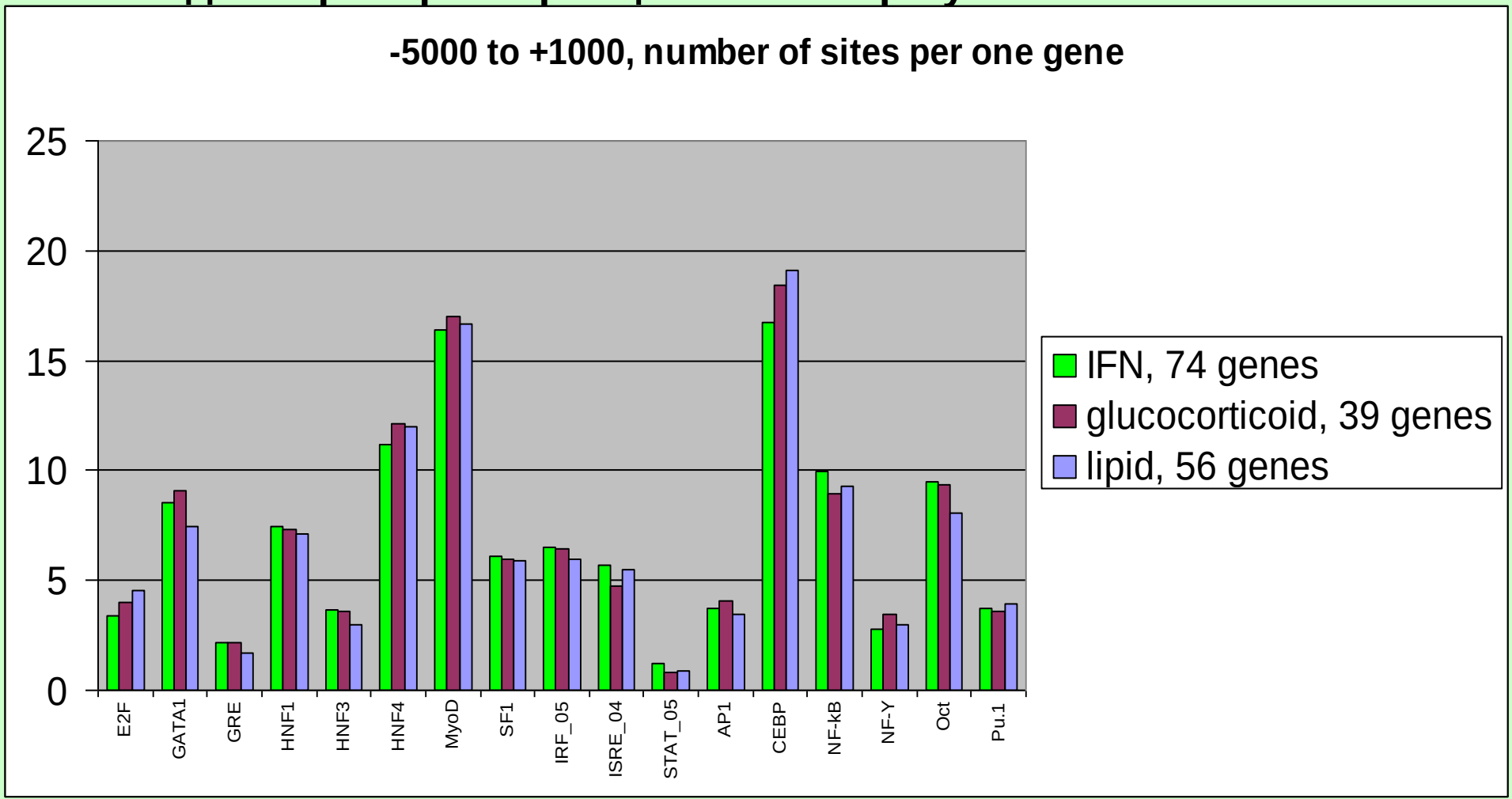
Определение характерных для интерферон-индуцируемых генов закономерностей в расположении сайтов связывания разных транскрипционных факторов



Задача 6

Определение характерных для интерферон-индуцируемых генов закономерностей в расположении сайтов связывания разных транскрипционных факторов

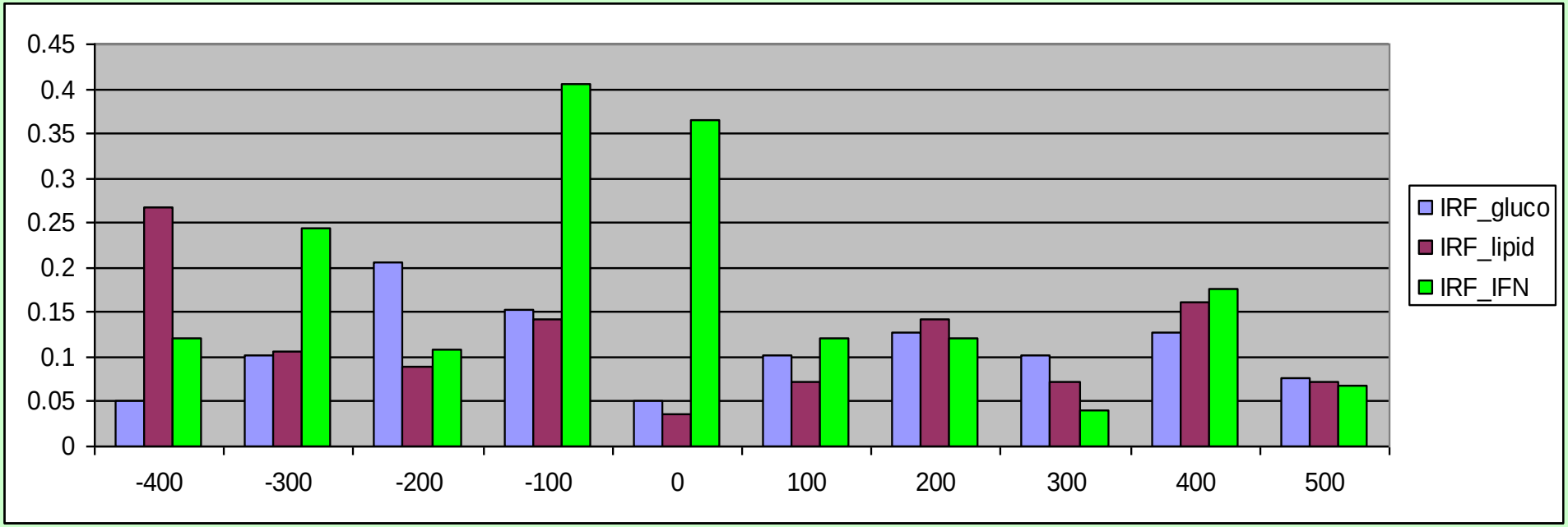
Выборка из последовательностей 74 генов человека, по 5 000 п.о. до старта транскрипции и после poly-A сайта.



Задача 6

Определение характерных для интерферон-индуцируемых генов закономерностей в расположении сайтов связывания разных транскрипционных факторов

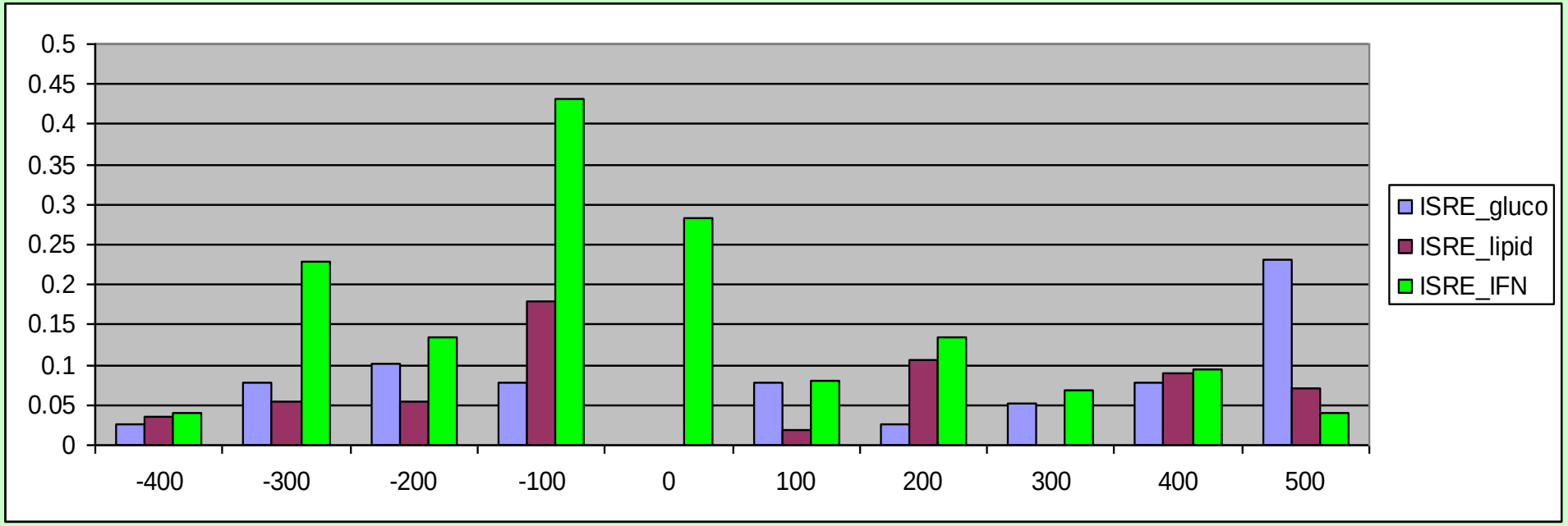
Распределение потенциальных сайтов связывания IRF-1 в промоторных районах различных функциональных групп генов



Задача 6

Определение характерных для интерферон-индуцируемых генов закономерностей в расположении сайтов связывания разных транскрипционных факторов

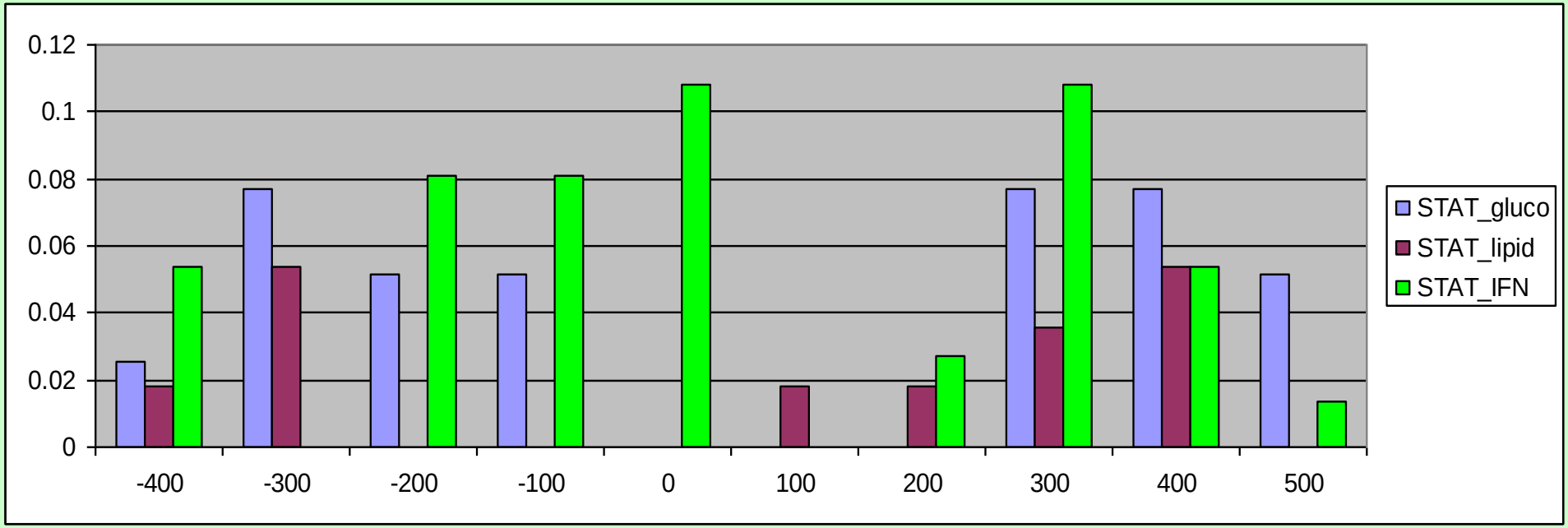
Распределение потенциальных сайтов связывания ISGF3 в промоторных районах различных функциональных групп генов



Задача 6

Определение характерных для интерферон-индуцируемых генов закономерностей в расположении сайтов связывания разных транскрипционных факторов

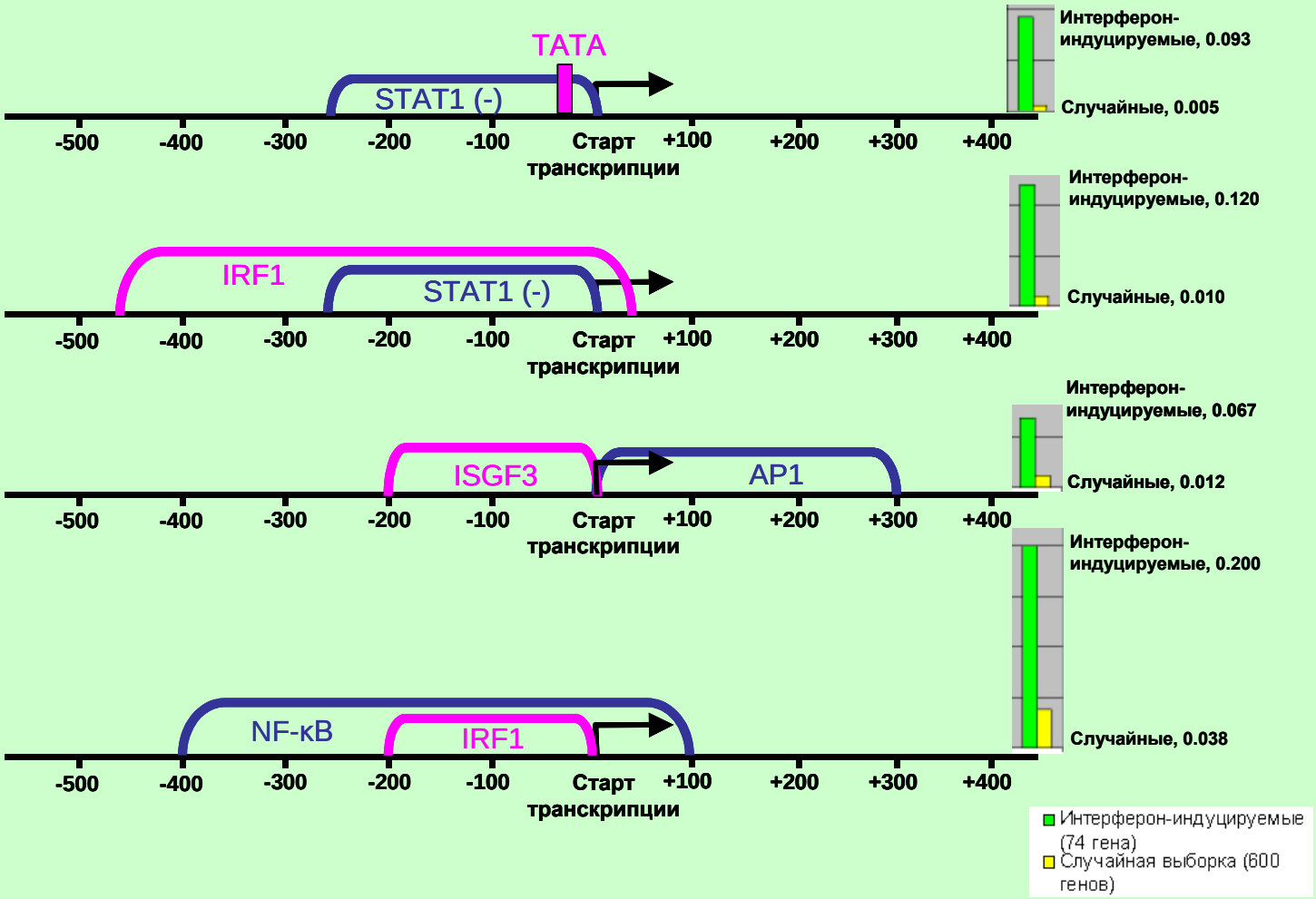
Распределение потенциальных сайтов связывания STAT1 в промоторных районах различных функциональных групп генов



Задача 6

Определение характерных для интерферон-индуцируемых генов закономерностей в расположении сайтов связывания разных транскрипционных факторов

Частота встречаемости различных пар сайтов



Задача 7

Разработка методов распознавания интерферон-индуцируемых промоторов и энхансеров в геномах эукариот

Проанализировано несколько сотен комбинаций сайтов, из которых отобрано 159, имеющих статистически значимое отличие в частоте встречаемости у исследуемых генов по отношению к контрольной выборке (EPD, -1000 +1000; 1664 последовательности промоторов человека)

Метод	Кол-во используемых комбинаций
Метод 0 - распознавание любых интерферон-индуцируемых генов	28
Метод 1 - распознавание генов, индуцируемых интерферонами I типа (IFNα, IFNβ)	23
Метод 2 - распознавание генов, индуцируемых интерферонами II типа (IFNγ)	18

Задача 7

Разработка методов распознавания интерферон-индуцируемых промоторов и энхансеров в геномах эукариот

Для проверки разработанных методов создана база данных ИИГ по опубликованным результатам анализа с помощью РНК микрочипов. База содержит данные о времени, типе и уровне индукции, клетках, в которых был проведен эксперимент, а также последовательности ДНК для 1005 генов.

Выборка	Количество последовательностей в выборке
Выборка M0 (все ИИГ, идентифицированные с помощью микрочипового анализа)	1005
Подвыборка M1 (только гены, индуцируемые интерферонами первого типа - ИФα, ИФβ)	668
Подвыборка M2 (только гены, индуцируемые интерфероном второго типа, ИФγ)	97

Задача 7

Разработка методов распознавания интерферон-индуцируемых промоторов и энхансеров в геномах эукариот

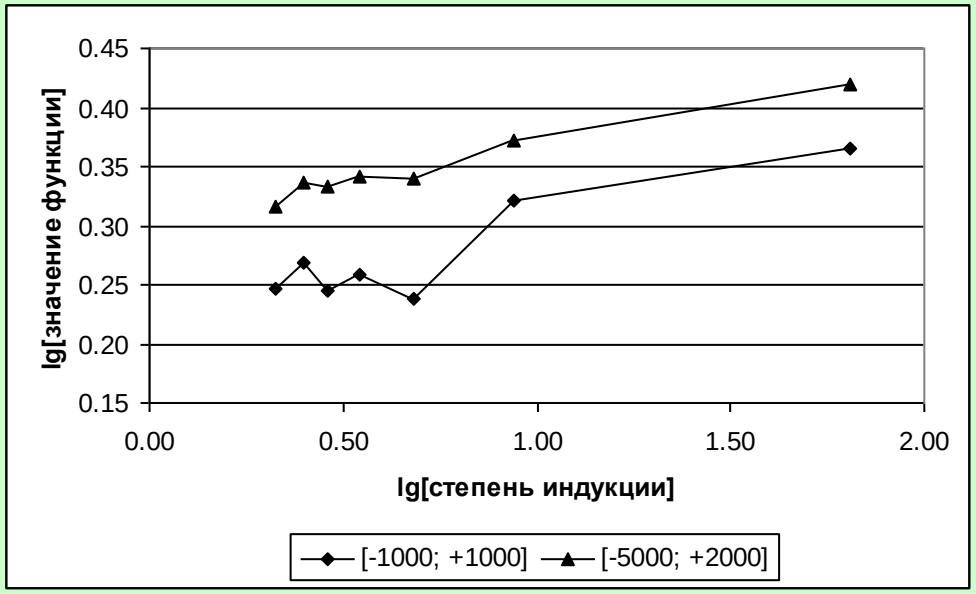
Показано, что выборки генов, полученные с помощью данных микрочипового анализа, действительно обогащены ИИГ

Выборка	Общее количество последовательностей в выборке	Распознано генов	Распознано в %
Обучающая выборка	72	17	23.6
Выборка по данным микрочипового анализа (M0)	1005	156	15.5
EPD	1664	74	4.4
Глюкокортикоид-регулируемые гены	70	0	0
Гены липидного метаболизма	58	0	0

Задача 7

Разработка методов распознавания интерферон-индуцируемых промоторов и энхансеров в геномах эукариот

Полученные выборки и подвыборки проверены с помощью созданных методов распознавания ИИГ. Установлено, что значения функций распознавания коррелируют с уровнем индукции генов.



Распознавание районов, отвечающих на индукцию любыми интерферонами (метод 0) по неполной выборке M0 (808 генов из 1005, для которых были количественные данные по степени индукции). Зависимость логарифмических значений функции распознавания и степени индукции при использовании более длинных (от -5000 до +2000 относительно старта транскрипции) и более коротких (от -1000 до +1000 относительно старта транскрипции) последовательностей.

Для оценки статистической значимости использовался стандартный критерий χ-квадрат. Для этого критерия использовался уровень значимости p-value=0.01.

Задача 8

Поиск потенциальных интерферон-индуцируемых генов человека

Изучено промоторных районов человека (-1000; +1000)	1 664
Распознано интерферон-индуцируемых генов	63
Из них присутствует в обучающей выборке	19
Новых	44

Задача 8

Поиск потенциальных интерферон-индуцируемых генов человека

Примеры генов, которые были распознаны как интерферон-индуцируемые:

- **TDO2 (tryptophan 2,3-dioxygenase)**
- **RPS27 (ribosomal protein S27 / metallopanstimulin 1)**

были получены данные microarray, подтверждающие наши предсказания

- **M6PR (cation dependent mannose-6-phosphate receptor)**

был экспериментально показан усиленный синтез белка, кодируемого этим геном, в клетках дыхательного эпителия под действием интерферона-γ

- **PIP (gross cystic disease fluid protein)**
- **RABAC1 (Prenylated Rab acceptor protein 1)**

подтверждения регуляции интерферонами найдено не было, однако, по свойствам и функциям они очень похожи на некоторые интерферон-индуцируемые гены, присутствующие в нашей контрольной выборке

Выводы:

1. 1. Разработана компьютерная технология GeneNet, предназначенная для реконструкции генных сетей про- и эукариот на основе аннотации экспериментальных данных из научных публикаций, а также для их автоматической визуализации и анализа структурно-функциональной организации. С использованием технологии GeneNet в ИЦиГ СО РАН на основе аннотации 7755 научных публикаций осуществлена реконструкция структурно-функциональной организации около 100 генных сетей про- и эукариот, включающих описание 2112 генов, 3711 белков, более 11500 взаимодействий между различными компонентами.

2. 2. С использованием компьютерной технологии GeneNet реконструированы генные сети интерфероновой индукции у млекопитающих, включающие более 100 генов, 200 белков и 500 молекулярных взаимодействий. Проведен компьютерной анализ топологии этих сетей. Показано преобладание регуляторных контуров с положительными обратными связями (88 против 14 с отрицательными), обеспечивающее усиление ответа клетки на интерфероны.

3. 3. На основе компьютерной аннотации экспериментальных данных из научных публикаций в базе данных IIG-TRRD аннотировано 238 протяженных регуляторных районов интерферон-индуцируемых генов млекопитающих (130 генов, 666 сайтов, 752 публикации).

4. На основе аннотированных в IIG-TRRD сайтов связывания транскрипционных факторов ISGF3, IRF1, STAT1, NF-κB, AP1, которые играют важную роль в функционировании генных сетей интерфероновой индукции, созданы методы распознавания этих типов сайтов.

Выводы:

1. 5. Проведен сравнительный анализ расположения потенциальных сайтов связывания транскрипционных факторов в регуляторных районах генов липидного метаболизма, клюкокортикоид-регулируемых генов и интерферон-индуцируемых генов. Показано, что плотность сайтов связывания транскрипционных факторов IRF1, ISGF3, STAT1, NF-κB, в районе [-200; +1] относительно старта транскрипции в интерферон-индуцируемых генах достоверно выше, чем в двух других группах генов.

6. Найдены специфические для ИИГ закономерности расположения пар сайтов связывания транскрипционных факторов. На основе этих закономерностей построено три метода распознавания ИФ-индуцируемых районов ДНК: (1) индуцируемых любыми ИФ; (2) индуцируемых ИФ-α и ИФ-β; (3) индуцируемых ИФ-γ.

7. Создана база данных ИИГ по опубликованным результатам анализа с помощью РНК микрочипов. База содержит данные о времени, типе и степени индукции, клетках, в которых был проведен эксперимент, а также последовательности ДНК для 1005 генов. Гены, собранные в этой базе, проверены с помощью созданных методов распознавания ИИГ. Установлено, что выборки генов, полученные с помощью данных микрочипового анализа, действительно обогащены ИИГ, и значения функций распознавания коррелируют с уровнем индукции генов.

4. 8. Проведен поиск потенциальных генов-мишеней интерфероновой индукции в промоторных районов генов человека, экстрагированных из базы данных EPD. Найдено 74 гена, с высокой вероятностью индуцирующихся интерферонами. Сделана приблизительная оценка количества интерферон-индуцируемых генов в геноме человека, согласно которой общее количество таких генов превышает 3000.

Публикации:

По теме диссертации опубликовано

90 работ,

из них 31 в рецензируемых журналах,

имеются свидетельства об официальной регистрации двух баз данных,

поддерживается три web-сайта