Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего профессионального образования

Национальный исследовательский университет
«Высшая школа экономики»

Факультет бизнеса и менеджмента
Школа бизнес-информатики
Кафедра инноваций и бизнеса в сфере информационных технологий

Выпускная квалификационная работа на тему

«Анализ рынка и создание фреймворка для определения перспективности биоинформатического стартапа на основе баз данных»

Студент группы ББИ-122 Намазов Р.К.

Руководитель ВКР д.ф.-м.н., проф. Ройтберг М.А.

Ст. преподаватель

Москва – 2016

Оглавление

Введение

С начала проекта «Геном человека»[1] мир биоинформатики начал расти поразительными темпами. Количество данных, которые появились после этого исследования колоссально. Самым логичным шагом после сбора такого количества информации, является создание методов для хранения и анализа этих данных. В результате появилось большое количество проектов, главная цель которых хранить биологические данные в общедоступном формате. Наиболее примечательное и успешное начинание это PDB[2]. Protein Data Bank - это ресурс, который дает ученым по всему миру возможность узнавать о биологических структурах, которые были исследованы другими учеными, а также загружать результаты своих исследований.

Со временем стало понятно, что у появляющихся средств для хранения биологических данных есть очевидные недостатки. Одним из таких недостатков, который является критичным для статистических исследований в области биологических структур, является избыточность. Для таких «популярных» структур, как белки существуют механизмы для отсеивания избыточности. Однако для структур РНК нет надежных механизмов, которые позволяют использовать безубыточные техники для поиска.

Несмотря на то, что способы определения 3D структуры РНК значительно продвинулись, данную процедуру нельзя назвать рутинной. Это привело к ситуации, когда любое статистическое исследование, основанное на РНК структурах, страдает от избыточности.

Наличие спроса на биоинформационные технологии и рост рынка, который будет показан в основной части данной работы, привело к огромному количеству проектов в данной сфере. Однако такая наукоемкая сфера, как биоинформатика требует тщательного анализа рынка и просчета потенциальной перспективности для того, чтобы максимизировать потенциальные инвестиции, которые может получить молодая компания, работающая в этой сфере. В данной работе предпринята попытка проанализировать рынок биотехнологий, а также показать, как реальный наукоёмкий стартап в него вписывается.

---

[1] Human Genome Project Information Archive: 1990-2003 // U.S. DOE Human Genome Project. Дата обновления: 09.05.2016. URL: http://web.ornl.gov/sci/techresources/Human_Genome/ (дата обращения: 10.05.2016).

[2] The Protein Data Bank: [Электронный ресурс] // Research Collaboratory for Structural Bioinformatics. 2003. URL: http://www.rcsb.org/pdb/home/home.do (дата обращения: 15.04.2016).

Как будет показано в основной части, наукоемкие стартапы отличаются от классических бизнесов и стартапов тем, что инновации, потенциальная научная польза и новизна превалируют. Однако, выделение фреймворка после определения идеи и рынка является хорошим механизмом для потенциальной оценки стартапа. Инновационная составляющая делает поэтапную схему в следующем порядке предпочтительно:

- Анализ рынка и определение потребности
- Формулирование и реализация технической идеи (MVP)[3]
- Оценка и измерение перспективности стартапа с помощью фреймворка построенного под данный стартап.

Главную *проблему,* которую пытается решить данная работа можно сформулировать следующим образом: отсутствие средств для проведения исследований на неизбыточном множестве РНК структур привело к необходимости создать соответствующего проекта. Создание такого наукоемкого стартапа не может происходить без анализа рынка и попытки оценки перспективности проекта.

Цель исследования можно определить как попытку построить соответствующий наукоёмкий стартап, который дает возможность ученым пользоваться неизбыточными базами данных РНК структур. Анализ рынка выступает в качестве одно из критериев, который показывает целесообразность такого проекта, а также выступает в качестве информационной сводки о состоянии рынка биоинофрмационных стартапов. С помощью фреймворка, который будет описан в практической части работы, мы пытаемся показать значимость данного проекта, а также оценить его потенциальную перспективность.

Целесообразно сформулировать *задачи* данной работы. План работы выглядит следующим образом:

- Проанализировать рынок биоинформационных стартапов, это включает в себя:
    - Описание рынка
    - Обзор стартапов, представленных на рынке
    - Прогнозы роста рынка
- Показать устройство стартапа на основе неизбыточный базы данных РНК структур, подразумеваются следующие подзадачи:
    - Описать теоретический базис построение неизбыточный базы данных РНК структур

[3] Minimal viable product. Минимальная версия продукта, которая отражает основной функционал

- Показать программные и алгоритмические методы, которые был проведены для построения
- Обзор функционал существующей базы
- Описать критерии, составляющие фреймворк для оценки перспективности

Исходя из поставленных задач текст данной практической исследовательской работы будет построен следующим образом.
В начале основной части будет описан рынок биоинформационных стартапов и проанализированы успешные стартапы последних лет. Будет сделан упор на успешные биоинформационные стартапы, которые получили значительные венчурные инвестиции.

Далее будут описаны теоретические моменты, связанные с построением неизбыточный базы данных РНК структур, а также описаны работы в данной области. Теоретическая глава будет включать в себя краткое введение в предметную область с объяснением основных научных методов, которые используются в данной области.

Практическая часть будет посвящена нашим прикладным исследованиям в данной области, а также описанию программных решений, которые способствовали в построении неизбыточной базы данных. Будут описаны действия, которые были выполнены при попытке построения кластеров структур для последующего составления неизбыточный базы данных. Далее будут показаны прикладные программные инструменты, которые были использованы для построения базы данных, а также для анализа входных данных.

В заключительной главе основной части будут описаны критерии для составления фреймворка, который будет использован в попытке предсказания перспективности описанного наукоемкого стартапа.

В заключении будут подводиться итоги того, что было исследовано в работе. Также будут описаны предложения для будущих исследований.

Глава 1. Теоретическая часть

В основной части будет краткий обзор существующих решений и практик в области NR баз данных, а также будет описана предметная область, с которой работают NR базы данных. Помимо этого, будут очерчены границы рынка биотехнологических стартапов. Затем, будет кратко описан фреймворк для последующего подробного изучения. В конце первой главы будет обоснована структура работы.

Первым шагом в создании любого наукоемкого проекта является определение рынка. Компьютеры стали неотъемлемой частью молекулярной биологии после того, как Фредерик Сенгер определил последовательность инсулина в 1950 г.[4] Ручное сравнивание двух последовательностей оказалось непрактичным, поэтому позже директором национального центра биотехнологической информации, доктором Маргарет Дейхов была скомпилирована первая база данных последовательностей протеинов, а также были созданы первые методы для выравнивания последовательностей[5]. Со временем количество проектов, занимающихся биоинформатикой, стало расти поразительными темпами. Каждая крупная биологическая лаборатория имела штатного биоинформатика.

Вследствие высокой стоимости R&D, а также низкой прибыли в годы разработки, в индустрии доминируют большие компании. Это приводит к тому, что малый бизнес должен иметь существенные научные наработки и практические аппликации, чтобы котироваться на этом рынке. Поэтому рынок и идея (технологическая «изюминка») диктуют то, каким должен быть фреймворк для определения перспективности биоинформационного стартапа.

Предполагается, что рынок биоинформационных технологий вырастет до $12.86 млрд. к 2020 году с CAGR[6] в 21.2% к 2014-2020 (Grand View Research, 2015)[7]. Данный рынок можно сегментировать на три больших группы:

---

[4] Sanger F. The free amino groups of insulin // Biochemical Journal. 1945. Vol. 39. № 5. PP. 507-515. doi:10.1042/bj0390507

[5] Margaret Oakley Dayhoff 1925–1983 // Bulletin of Mathematical Biology. 1984. Vol. 46. Issue 4. PP. 467-472.

[6] Compound Annual Growth Rate

[7] Bioinformatics Market Analysis By Product (Sequence Analysis, Manipulation, Alignment, Structural & Functional Analysis Platforms, Data Management, Sequencing, Data Analysis Service, Generalized, Specialized Biocontent), By Application (Genomics, Molecular Phylogenetics, Metabolomics, Proteomics, Transcriptomics, Chemoinformatics & Drug Designing) And Segment Forecasts To 2020: report summary // Grand View Research, Inc. 2015.

1. средства для управления знаниями
2. платформы для биоинформатики
3. сервисы для биоинформатики

Рынок биоифнормационных платформ является самым прибыльным на конец 2013 года[8]. Это обусловлено повышенным использованием данных платформ в различных проектах, связанных с геномикой. Однако, ожидается, что средства для управления знаниями будут самым прибыльным сегментом вследствие с увеличенным количеством данных, появившихся после увеличенного количества клинических экспериментов.

На графике ниже видно предсказание биоинформационного рынка в Европе на период с 2012-2020 (в миллиардах долларов)[9]:

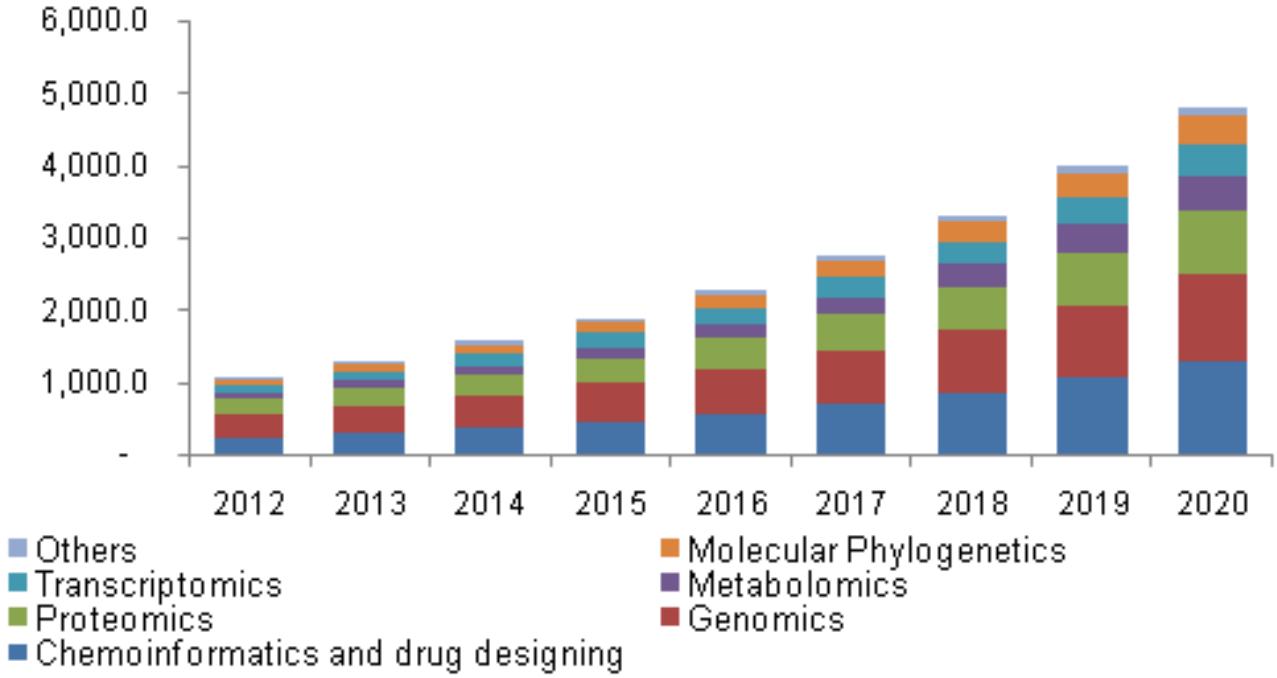


Рисунок №1

August. URL: http://www.grandviewresearch.com/industry-analysis/bioinformatics-industry (дата обращения: 18.04.2016).

[8] Reed J. Trends in Commercial Bioinformatics // Oscar Gruss Biotechnology review. New York, 2000. №13.

[9] Bioinformatics Market Analysis By Product … // Grand View Research, Inc. 2015. August. URL: http://www.grandviewresearch.com/industry-analysis/bioinformatics-industry (дата обращения: 05.05.2016).

Говоря о рынке биоинформационных стартапов следует отметить стартап компании, которые получили большое финансирование, это поможет показать потенциал рынка и его отношение к новым компаниям. Стоит отметить, что за последние годы количество венчурных сделок, связанных с биотехнологическими компаниями значительно увеличилось. Проведем краткий обзор крупных стартапов в этой области.

- 23andMe. Компания занимающяяся персональной геномикой. Она продает генетические тесты на прямую людям. Любой человек может заказать специальный набор, собрать биоматериал, отправить набор обратно, для того чтобы провели все необходимые исследования. Доступ к результатам исследования появится в лично кабинете пользователя. 23andMe [10] получили венчурные инвестиции в размере более $241 миллиона
- DNAnexus [11] . Компания, которая стремится стать ДНК платформой будущего. По заявлению компании, благодаря усовершенствованному оборудованию стоимость ДНК секвенсирования улучшается десятикратно каждые 18 месяцев. Как следствие, главной сложностью становится управление данными, которые появляются с огромной скоростью. DNAnexus предлагает инфраструктуру для управления данными необходимого масштаба. Объем привлеченных венчурных инвестиций достигает $30 миллионов.
- Transcriptic [12] . Данная компания создает облачную лабораторию для проведения экспериментов удаленно. Их главные достоинства это скорость проведения экспериментов, а также дешевизна и высокая степень повторяемости. Они убирают надобность в собственной лаборатории, что приводит к экономии денег и времени, которые раньше уходили на перенос маленьких объемов жидкости в ручную. Объем привлеченных инвестиций $20 миллионов.
- Soylent[13]. Относительно молодая компания, которая занимается, тем что создает еду в виде пудры, которая содержит в себе все необходимые ингредиенты для жизнедеятельности человека. Основной упор делается на полезность, дешевизну и скорость приготовления еды. Объем инвестиций более $21 миллиона.

[10] 23andMe. 2016. URL: https://www.23andme.com/en-int/ (дата обращения: 18.04.2016).
[11] DNAnexus. 2015. URL: https://www.dnanexus.com/ (дата обращения: 25.04.2016).
[12] Transcriptic. 2015. URL: https://www.transcriptic.com/ (дата обращения: 24.04.2016).
[13] Soylent. 2016. URL: https://www.soylent.com/ (дата обращения: 24.04.2016).

- Science Exchange[14] - это маркетплейс, которы предоставляет возможность исследователям заказывать эксперименты в лучших лабораториях мира. Их главная цель - улучшить качество и эффективность научных исследований с помощью создания удобной платформы, на которой ученые могут взаимодействовать. Объем инвестиций более $30 миллионов.
- Notable Labs [15] . Данная компания предоставляет персонализированные комбинационные тесты различных лекарств для пациентов больных раком. Каждая раковая опухоль - уникальное заболевание для каждого пациента, и каждый пациент по-разному реагирует на различные комбинации лекарств. Notable Labs позволяет подбирать наилучшее лечение для больных раком.
- Benchling[16] - платформа для удобной работы в сфере биотехнологий. Объем инвестиций более $5 миллионов.
- Cofactor Genomics[17]. Компания использующая РНК для диагностирования заболеваний.

Эти успешные стартапы всего лишь малая часть растущего рынка. Однако, этого достаточно, чтобы показать намерения венчурных инвесторов. Одна главная черта, которая объединяет эти стартапы - это желание сочетать передовые IT технологии с новейшими разработками в области биотехнологий. Каждый из этих стартапов отталкивается от идеи.

Рынок. Существующие подходы

Множество ученых по всему миру исследуют и открывают практически идентичные РНК структуры, которые отличаются в деталях несущественных для статистического исследования.

Одной из наиболее подробных работ про NR базы данных, является работа Neocles B. Leontis & Craig L. Zirbel «Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking»[18]

---

[14] Science Exchange. 2016. URL: https://www.scienceexchange.com/ (дата обращения:24.04.2016).

[15] Notable Labs. 2016. URL: https://notablelabs.com/ (дата обращения:24.04.2016).

[16]Benchling. 2015. URL: https://benchling.com/ (дата обращения:24.04.2016).

[17] Cofactor Genomics. 2013. URL: https://cofactorgenomics.com/ (дата обращения:24.04.2016).

[18] Leontis N.B. Zibrel C.L. Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking // Nucleic Acids and Molecular Biology. New York, 2012. Vol. 27. PP. 281-298.

Обзор существующих решений будет в большей степени построен на основе работы Leontis & Zirbel.

По состоянию на 2016 год, существует более 2 тысяч 3D структур РНК в PDB. Большее количество этих структур получено с использованием X-ray кристаллографии. Значительная часть данных структур статистически избыточна.

Построение NR базы данных биологических структур невозможно без следующих шагов:

- Определение избыточности внутри заданного PDB файла[19]
- Определение избыточности среди различных PDB файлов
- Выбор представителя кластера
- Взаимодействие с PDB. Автоматическое получение данных и обновление данных при обновлении информации.

Избыточность в файле

Прежде всего, чтобы разобраться с избыточностью внутри файла, надо понимать, вследствие чего она может возникать. Для этого необходимо понимать основы X-ray кристаллографии. Основные идеи заключаются в терминах «асимметричной единицы», «юнит ячейка» и «биологической единицы». Асимметричная единица содержит уникальную часть структуры кристалла. Наименьшую часть кристалла, из которой с помощью операций симметрии можно воссоздать «юнит ячейку». «Юнит ячейка» (unit cell) это повторяющаяся часть кристалла, с помощью, которой можно воссоздать полноценный кристалл.

[19] Файл, в котором хранится РНК структура

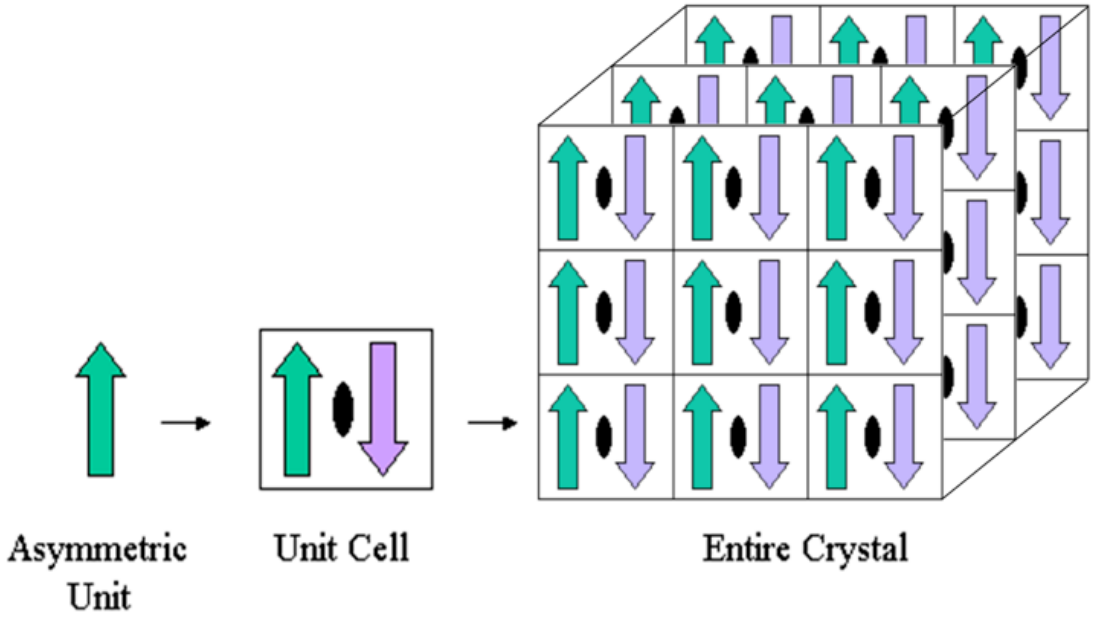


Рисунок №2

Биологическая единица — это структура, которая считается функциональной частью макромолекулы и обычно считается самой интересной частью. Файл со структурой кристалла после X-ray кристаллографии содержит одну асимметричную единицу. Обычно для анализа используют файлы с расширением «.pdb». В зависимости от позиции и конфигурации кристаллизованной макромолекулы в внутри «юнит ячейки» ASU (asymmetric unit) может содержать (1) часть функциональной единицы (2) одну полноценную биологическую единицу (3) несколько биологических единиц

Данный вид избыточности возникает вследствие того, что в файле PDB при X-ray кристаллографии находятся данные про биологическую единицу и про асимметричные единицы.

Избыточность между файлами

Есть множество причин почему избыточность между файлами существует.

- Много ученых исследует практически одинаковые структуры

- После того как «дикая» структура молекулы определена, ученые определяют различные свойства измененных молекул. Эти исследования загружаются в PDB, как новая структура

- Часто бывает необходимо определить механизмы молекулы в «действии». Это значит, например, получение снэпшота энзима во время реакции

- Существует большое число гомогенных молекул в PDB. Это структуры, которые эволюционно очень похожи и имею практически идентичные функциональные качества, но имеют различия из -за эволюционных мутаций

- Молекула может быть кристаллизована в различных формах

Как видно, существует большое количество случаев, когда проявляется избыточность между файлами. Наша главная цель это разбить различающиеся молекулы на разные классы и сопоставить похожие в одинаковые классы.

Выбор представителя

Далее следует процесс кластеризации, который заключается в распределении структур по классам эквивалентности таким образом, чтобы похожие классы были «близко», а различные «далеко». Кластеризация будет подробнее описана в главе «Наше решение».

Выбор представителя - важный этап, потому что от него зависит качество выборки, которая будет доступна конечному пользователю.
Предположительно, будут использоваться следующие критерии для определения представителя каждого класса эквивалентности:

- Количество аннотированных спаренных оснований
- Разрешение каждой структуры; чем меньше, тем лучше
- Количество неразрешенных биологических единиц; чем меньше, тем лучше
- Дата публикации структуры на PDB; новее - лучше

Первый критерий - наиболее важный, потому что он демонстрирует качество структуры и предоставляет наибольшее количество биологически-релевантной информации. Все спорные ситуации разрешаются следующими критериями в указанном порядке.

Наше решение

Одним из ключевых моментов в создании пополняемой NR базы данных является построение механизмов получения данных из общедоступных источников. Как мировую базу данных структур РНК, которую постоянно пополняют ученые, мы используем Protein Data Bank. Есть два основных момента во взаимодействии с PDB:

- Автоматическая загрузка и обновление данных из PDB
- Взаимодействие с данными в удобной манере.

Мы используем скрипты, написанные на Python для автоматической загрузки и обновления данных из PDB. Информация сохраняется в базу данных. Существует веб интерфейс для удобного взаимодействия с базой данных.

Разработанная процедура

Одна из проблем, которая была описана в существующих решениях, это проблема избыточности между файлами. Наш подход убирает данный вид избыточности за счет следующих действий:

- использования формата mmCIF вместо стандартного
- концентрация на полной структуре молекулы

Использование формата mmCIF позволяет полностью избежать проблемы с использованием асимметричной единицы, так как данный формат может содержать в себе всю структуру.

Предпосылками для создания финальной процедуры были следующие моменты:

1. Создание «наивной» кластеризации
2. Создание скриптов для анализа входных данных

Наивная кластеризация и новое решение

Гипотеза с наивной кластеризацией заключалась в следующем. Если использовать готовые инструменты для сравнения структур и объединить это с алгоритмом, основанным на предположении, что отношение избыточности является отношением эквивалентности.

Алгоритм можно описать следующим образом:

1. В начале имеется 0 кластеров и N последовательностей

2. При подаче первой последовательности получаем кластер из одной последовательности
3. При подаче следующей последовательности нужно определить дальнейшие действия: а) создать новый кластер или б) добавить её к существующему кластеру

Основным инструментом для сравнения последовательностей является библиотека BioPython[20].

Исходный код алгоритма можно увидеть в приложение под номером 1.

Данный способ разделения на классы не подходит, потому что сравнение больших структур, комбинированное с наивным подходом, дает очень большое время работы, которое невозможно на больших объемах данных.

Для создания более совершенного алгоритма кластеризации было необходимо создать скрипты, которые позволяют оценивать и анализировать входные данные в удобном формате.

Основной задачей скрипта было сгруппировать данные об РНК структурах с PDB по определенным критериям для последующего удобного анализа. Входные данные выглядели следующим образом:

[20] Biopython: freely available Python tools for computational molecular biology and bioinformatics

ID	PDB_ID	Length	Chains	Sequence	DBN	Molecules	Organisms	Fragments	Synthesized	EC	Engineered	Mutations	Details		
1	124D_1	16	A,B	DG,DT,DC,DA,DC,DA,DT,DG,C,A,U,G,U,G,A,C	((((((((((((((()))))))))))))))	DNA (5'-D("GP*TP*CP*AP*CP*AP*TP*G)-3') && RNA (5'-R("CP*AP*UP*GP*UP*GP*AP*C)-3')	None && None	None && None	None && None	None && None	YES && YES	None && None	CHEMICALLY SYNTHESIZED && CHEMICALLY SYNTHESIZED		
2	157D_1	24	A,B	C,G,C,G,A,A,U,U,A,G,C,G,C,G,C,G,A,A,U,U,A,G,C,G	(((.((((.((((.)))).)))))))	RNA (5'-R("CP*GP*CP*GP*AP*AP*UP*UP*AP*GP*CP*G)-3') && RNA (5'-R("CP*GP*CP*GP*AP*AP*UP*UP*AP*GP*CP*G)-3')	None && None	None && None	None && None	None && None	YES && YES	None && None	None && None		
3	165D_1	18	A,B	G,C,U,U,C,G,G,C,BRU,G,C,U,U,C,G,G,C,BRU	(((..(((.)))..))).	DNA/RNA (5'-R("GP*CP*UP*UP*CP*GP*GP*CP*)-D("(BRU)")-3') && DNA/RNA (5'-R("GP*CP*UP*UP*CP*GP*GP*CP*)-D("(BRU)")-3')	None && None	None && None	None && None	None && None	YES && YES	None && None	None && None		
4	176D_1	12	A,B	GPN,APN,APN,CPN,TPN,CPN,G,A,G,U,U,C	(((((((((())))))))))	DNA (5'-D("GPN*APN*APN*CPN*TPN*CPN)-3') && RNA (5'-R(P*GP*AP*GP*UP*UP*C)-3')	None && None	None && None	None && None	None && None	YES && YES	None && None	CHEMICALLY SYNTHESIZED && CHEMICALLY SYNTHESIZED		
5	176D_10	12	A,B	GPN,APN,APN,CPN,TPN,CPN,G,A,G,U,U,C	(((((((((())))))))))	DNA (5'-D("GPN*APN*APN*CPN*TPN*CPN)-3') && RNA (5'-R(P*GP*AP*GP*UP*UP*C)-3')	None && None	None && None	None && None	None && None	YES && YES	None && None	CHEMICALLY SYNTHESIZED && CHEMICALLY SYNTHESIZED		
6	176D_2	12	A,B	GPN,APN,APN,CPN,TPN,CPN,G,A,G,U,U,C	(((((((((())))))))))	DNA (5'-D("GPN*APN*APN*CPN*TPN*CPN)-3') && RNA (5'-R(P*GP*AP*GP*UP*UP*C)-3')	None && None	None && None	None && None	None && None	YES && YES	None && None	CHEMICALLY SYNTHESIZED && CHEMICALLY SYNTHESIZED		
7	176D_3	12	A,B	GPN,APN,APN,CPN,TPN,CPN,G,A,G,U,U,C	(((((((((())))))))))	DNA (5'-D("GPN*APN*APN*CPN*TPN*CPN)-3') && RNA (5'-R(P*GP*AP*GP*UP*UP*C)-3')	None && None	None && None	None && None	None && None	YES && YES	None && None	CHEMICALLY SYNTHESIZED && CHEMICALLY SYNTHESIZED		
8	176D_4	12	A,B	GPN,APN,APN,CPN,TPN,CPN,G,A,G,U,U,C	(((((((((())))))))))	DNA (5'-D("GPN*APN*APN*CPN*TPN*CPN)-3') && RNA (5'-R(P*GP*AP*GP*UP*UP*C)-3')	None && None	None && None	None && None	None && None	YES && YES	None && None	CHEMICALLY SYNTHESIZED && CHEMICALLY SYNTHESIZED		
9	176D_5	12	A,B	GPN,APN,APN,CPN,TPN,CPN,G,A,G,U,U,C	(((((((((())))))))))	DNA (5'-D("GPN*APN*APN*CPN*TPN*CPN)-3') && RNA (5'-R(P*GP*AP*GP*UP*UP*C)-3')	None && None	None && None	None && None	None && None	YES && YES	None && None	CHEMICALLY SYNTHESIZED && CHEMICALLY SYNTHESIZED		
10	176D_6	12	A,B	GPN,APN,APN,CPN,TPN,CPN,G,A,G,U,U,C	(((((((((())))))))))	DNA (5'-D("GPN*APN*APN*CPN*TPN*CPN)-3') && RNA (5'-R(P*GP*AP*GP*UP*UP*C)-3')	None && None	None && None	None && None	None && None	YES && YES	None && None	CHEMICALLY SYNTHESIZED && CHEMICALLY SYNTHESIZED		
11	176D_7	12	A,B	GPN,APN,APN,CPN,TPN,CPN,G,A,G,U,U,C	(((((((((())))))))))	DNA (5'-D("GPN*APN*APN*CPN*TPN*CPN)-3') && RNA (5'-R(P*GP*AP*GP*UP*UP*C)-3')	None && None	None && None	None && None	None && None	YES && YES	None && None	CHEMICALLY SYNTHESIZED && CHEMICALLY SYNTHESIZED		
12	176D_8	12	A,B	GPN,APN,APN,CPN,TPN,CPN,G,A,G,U,U,C	(((((((((())))))))))	DNA (5'-D("GPN*APN*APN*CPN*TPN*CPN)-3') && RNA (5'-R(P*GP*AP*GP*UP*UP*C)-3')	None && None	None && None	None && None	None && None	YES && YES	None && None	CHEMICALLY SYNTHESIZED && CHEMICALLY SYNTHESIZED		
13	176D_9	12	A,B	GPN,APN,APN,CPN,TPN,CPN,G,A,G,U,U,C	(((((((((())))))))))	DNA (5'-D("GPN*APN*APN*CPN*TPN*CPN)-3') && RNA (5'-R(P*GP*AP*GP*UP*UP*C)-3')	None && None	None && None	None && None	None && None	YES && YES	None && None	CHEMICALLY SYNTHESIZED && CHEMICALLY SYNTHESIZED		
14	17RA_1	21	A	G,G,C,G,U,A,A,G,G,A,U,U,A,C,C,U,A,U,G,C,C	(((((..((.....)).)))))	RNA	None	RBS AND START SITE FOR PHAGE GA REPLICASE GENE	None	None	YES	A5U, A6U	IN VITRO SYNTHESIS FROM DNA TEMPLATE USING T7 RNA POLYMERASE. HAIRPIN CORRESPONDS TO NT -16 - +5 OF PHAGE GA REPLICASE AND THE YEAST PRE-MRNA BRANCHPOINT HELIX		

Рисунок №3

Итоговый скрипт имеет следующую функциональность:

1) Подсчет количества одинаковых организмов (сравнение идет только по первой структуре)
2) Подсчет количества одинаковых молекул. В случае, когда поле fragments не равно None&&None…, нужно сравнивать поле, составленное следующем образом. Если молекула имеет формат x && y && …. && z и фрагменты имеет формат n && m &&….&& l, то нужно строить строку вида: x—n && y—m &&…&& z—l
3) Подсчет количества одинаковых по молекулам и организмам одновременно
4) Подсчет одинаковых (по критериям выше) для первых моделей документов из PDB (структуры с PDB_Id имеющих _1 в тексте).

Вследствие несовершенства «наивного» алгоритма, с помощью скриптов для анализа входных данных была разработана более совершенная процедура:

1. На вход подается 0 кластеров и N последовательностей для кластеризации. Однако, теперь из них строится граф
2. Ребро к новой структуре добавляется только тем вершинам, у которых:
    1. тот же организм
    2. тот же класс типа молекулы (определяется по специальной функции)
    3. «похожи» последовательности

Степень похожести будет определяться следующим образом:

1. Сравниваем на точное совпадение. Если оно присутствует, то добавляем ребро, иначе шаг 2.
2. По первой последовательности создаем словарь её десяток. Затем, во второй последовательности берём каждую десятку и проверяем, есть ли она в словаре. Если есть, то по каждому индексу из списка в словаре выполняем следующие действия:
    1. Если предшествующие десятка нуклеотидов совпадает, то «забываем» про неё и идём дальше
    2. Иначе, пробуем максимально прожить совпадение вперёд и записываем его в виде тройки чисел x,y,z, где x- индекс начала нашего совпадения в 1-й последовательности, y - индекс начала совпадения во 2-й последовательности, z - длина совпадения. Из получившихся троек составляем «максимальную возрастающую последовательность» и считаем её общую длину M (без пересечений)

3. Сравниваем M / L1 с П1. L1 - длина большей последовательности, а П1 - порог 1. Если значение M/L1 >= П1 - добавляем ребро, иначе идем к шагу 4.

4. Сравниваем M / L2 с П2. Здесь L2 – длина меньшей последовательности, а П2 – порог 2. Если значение M/L2 >= П2 – добавляем ребро.

Следующий шаг — это оценка транзитивности. Делаем это следующим образом. Из полученного графа выделяются компоненты связанности. Оцениваем «долю транзитивности» - сколько ребер в каждой компоненте не хватает до клики, т.е. в каждой компоненте делим число ребер на N*(N-1)/2, где N - число вершин в компоненте.

Далее разбиваем граф на кластеры. Идем по всем ребрам и сравниваем для пары вершин их структуры. Если сравнение «неудачно», то удаляем ребро. В результате, компоненты связности в полученном графе являются нашими кластерами.

Монетизация проекта

Любой бизнес, а наукоёмкий стартап это потенциальный бизнес, должен приносить прибыль. Существует множество моделей монетизации стартапов, которые предоставляют услуги баз данных. Опишем основные модели по монетизации данных, а также покажем популярные схемы монетизации на примере крупных игроков.

Согласно исследованию компании Гартнер, методы монетизации данных делятся на два подхода[21]:

- Непрямая монетизация
    - Использование данных для улучшения собственных услуг
    - Использование данных для создания новых продуктов и рынков
    - Использование данных для укрепления партнерских отношений
- Прямая монетизация
    - Обмен или продажа данных и/или доступа к данным
    - Продажа подписок к исследованиям на основе данных

Также Roger Ehrenberg из IA Ventures определяет три основных типа компаний занимающихся продуктами связанными с данными[22]:

- *Содействующая база данных*. Данный тип баз данных используется, когда платформа просит пользователей внести данные, чтобы пополнить свою базу данных. Примером является PDB, который используется как основа для NR базы данных.
- *Платформы по процессингу данных*. Бизнесы, работающие по данной системе предоставляют пользователям доступ к данным в различных форматах, которые им нужны. В данный формат хорошо вписывается NR база данных
- *Платформы по созданию данных.* Данные сервисы специализируются на предоставлении механизмов для хранения и/или создания данных. Данные бизнесы ценны, тем что позволяют клиентам создавать собственные

[21] Methods for Monetizing Your Data [Электронный ресурс]: URL: http://www.gartner.com/webinar/3098518 (дата обращения: 05.05.2016).

[22] Ehrenberg R. Creating competitive advantage through data // IA Ventures' blog. 2011. URL: http://fortune.com/2011/07/21/creating-competitive-advantage-through-data/ (дата обращения: 10.05.2016).

решения на основе удобных механизмов, предоставляемых сторонней компанией.

Если говорить ближе к реальной практике, наиболее популярных механизмы монетизации, которые используются:

- *Оплата по запросам.* Данный способ монетизации заключается в оплате количества запросов, которые производятся к онлайн базе данных. Отличным примером является сервис от Microsoft. DynamoDB, который предоставляет определенное количество запросов бесплатно и делает услугу по обращению к своим серверам платной после определенного порога запросов. Данный способ меньше подходит к NR DB, так как из-за специфичности базы данных, ее использование будет заключаться не в постоянных запросах, а в работе с базой данных в определенных промежутках времени, которые привязаны к реальным исследованиям.
- *Месячная подписка*. Данный способ монетизации состоит в том, чтобы продавать подписки на использование сервисов, которые действуют месяц. Данную схему очень часто используют научные журналы. Технологическим примером является Heroku, компания, которая предоставляет услуги по размещению виртуальных серверов. Оплата производится по месячно.
- *Лицензия.* Продажа лицензии на безграничное использовании базы данных является одним из самых предпочтительных для научных сервисов. Высокая цена пожизненной лицензии является хорошим сигналом для клиентов и способствует в снижении морального риска между покупателем и продавцом.
- Open source + платная поддержка. Данный вид монетизации является предпочтительным, потому что показал себя наиболее эффективным в проектах, которые сделаны не огромными корпорациями. Отличными примерами являются компании, создавшие Redis, MongoDB и т.д. Основными причинами успешности данной схемы:
    - Решения, которые распространяются как свободное ПО получают больший отклик от сообщества. А также гораздо большее количество пользователей. Вследствие чего количество платных пользователей также увеличивается
    - Распространение по такой схеме делает возможным укрепится на рынке и получать поддержку от разработчиков со всего мира.

Глава 2. Практическая часть

Описание URSDB

URSDB это название базы данных, которая является бета-версией проекта. В базе содержится более 2935 РНК содержащих структур. 7718 цепей, 1314360 спаренных оснований различных типов и 5130 псевдоузлов.

URSDB - это релятивная база данных, работающая на базе MySQL. База состоит из большого количества таблиц разделенных на 4 группы:

1) таблицы данных, хранящихся в PDB (цепи, остатки, атомы и т.д.)
2) таблицы с данными из результатов DSSR[23] (спаренные основания, спирали и т.д.)
3) таблицы с данными о структурных мотивах, скомпилированное с использованием нашего программного пакета (треды, крылья, циклы, стемы, псевдоузлы и т.д.)
4) вспомогательные таблицы (параллельные стемы, РНК-протеиновые водородные связи)

[23] Программ для Dissecting the Spatial Structure of RNA

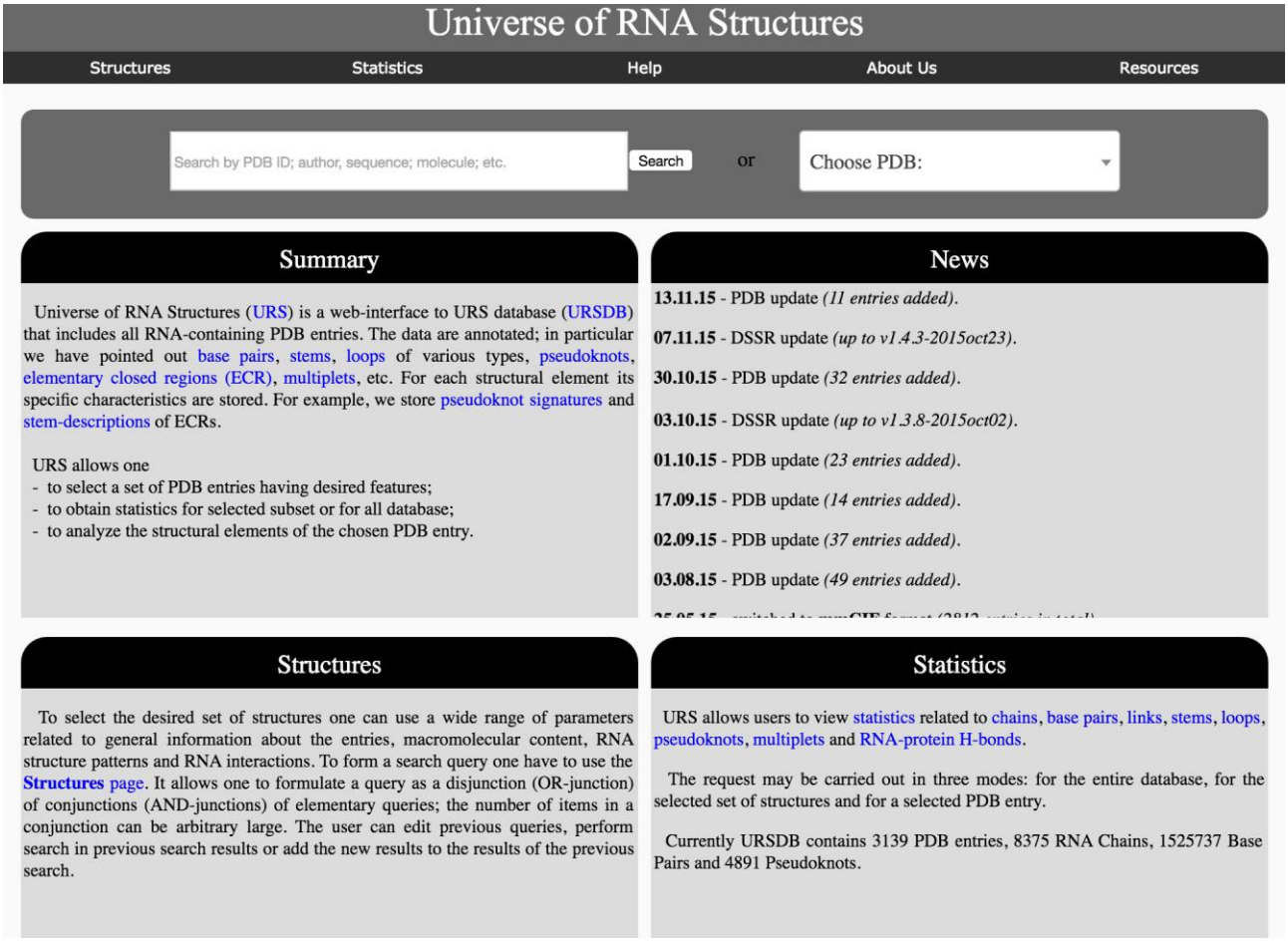


Рисунок №4

Главная страница состоит из следующих блоков:

- Меню с возможностью перехода в разделы:
    - поиска по структурам; многочисленные возможности поиска по различным параметрам
    - статистика; в данном разделе можно смотреть статистику, относящуюся к структурным элементам
    - помощь; подробная инструкция, о том, как пользоваться сайтом
    - раздел с информацией про авторов
    - раздел с ссылкой на используемые ресурсы
- Раздел для быстрого поиска по базе
- Блоки с кратким описанием основных разделов
- Блок с новостями

Раздел структуры

Time used: 4.09 sec

Query History Help

PDB has words: **124D**, **1**; SORT BY #Matches

Select Fields Help

N PDB ID | Header | Date | Method | Resolution

Sort by: # Matches | 1 to N (A to Z) | Show

Structures List CSV Export Help

Check All

Structures found: 3139

N	PDB ID (# Models)	Header	Date	Method	Resolution
1	124D (1)	DNA-RNA HYBRID	1993-05-07	SOLUTION NMR	
2	157D (1)	RNA	1994-02-01	X-RAY DIFFRACTION	1.8
3	165D (1)	DNA-RNA HYBRID	1994-03-21	X-RAY DIFFRACTION	1.55
4	176D (10)	PEPTIDE NUCLEIC ACID/RNA	1994-05-17	SOLUTION NMR	
5	17RA (12)	RNA	1998-08-04	SOLUTION NMR	
6	1A1T (25)	Viral protein/RNA	1997-12-15	SOLUTION NMR	
7	1A34 (1)	Virus/RNA	1998-01-28	X-RAY DIFFRACTION	1.81
8	1A3M (20)	RNA	1998-01-22	SOLUTION NMR	
9	1A4D (1)	RNA	1998-01-29	SOLUTION NMR	
10	1A4T (20)	TRANSCRIPTION/RNA	1998-02-04	SOLUTION NMR	
11	1A51 (9)	RNA	1998-02-19	SOLUTION NMR	
12	1A60 (24)	RNA	1998-03-04	SOLUTION NMR	
13	1A9L (12)	RNA	1998-04-07	SOLUTION NMR	
14	1A9N (1)	RNA BINDING PROTEIN/RNA	1998-04-08	X-RAY DIFFRACTION	2.38
15	1AC3 (8)	DNA/RNA HYBRID	1997-02-11	SOLUTION NMR	
16	1AFX (13)	RNA	1997-03-15	SOLUTION NMR	
17	1AJF (1)	RNA	1997-05-02	SOLUTION NMR	

Рисунок №5. Результаты поиска PDB структуры с ID 124D_1.

Результаты поиска по структурам состоят из трех блоков:

1. История поиска. В данном разделе показываются запросы, которые уже были сделаны
2. Раздел выбора полей. Здесь можно выбрать поля, которые будут показываться в регулирующей таблице, а также различные параметры для сортировки.
3. Результаты поиска, которые отформатированы согласно настройкам, указанным ранее. Также есть возможность экспорта результатов поиска в различные форматы (CSV, XLS, TXT, XML)

Поисковой интерфейс поддерживает большое количество поисковых запросов, относящихся к общей информации о:

- PDB документе
- молекуле

- РНК паттерне
- спаренных основаниях

РНК паттерн может быть описан с помощью фрагмента последовательности, dot-bracket нотации, сигнатуры псевдоузлов или ECR паттерна. Также у пользователя есть возможность указывать присутствие или отсутствие в структуре петлей различных типов, псевдо узлов и т.д.

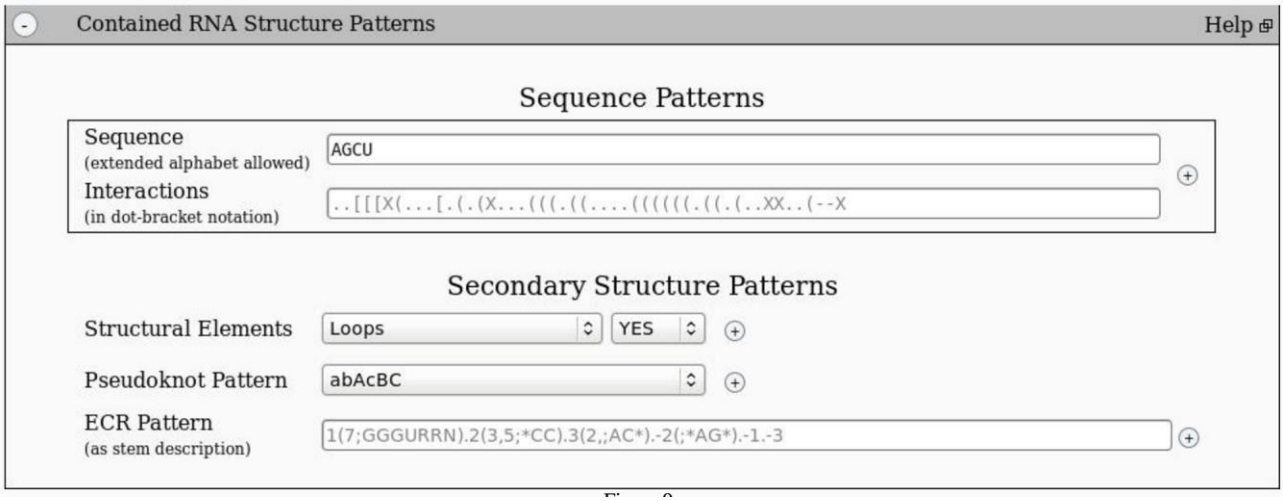


Рисунок №6

При нажатии на PDB_id можно посмотреть на подробно описание структуры, а также смоделированную 3D структуры, которую можно «крутить» в реальном времени.

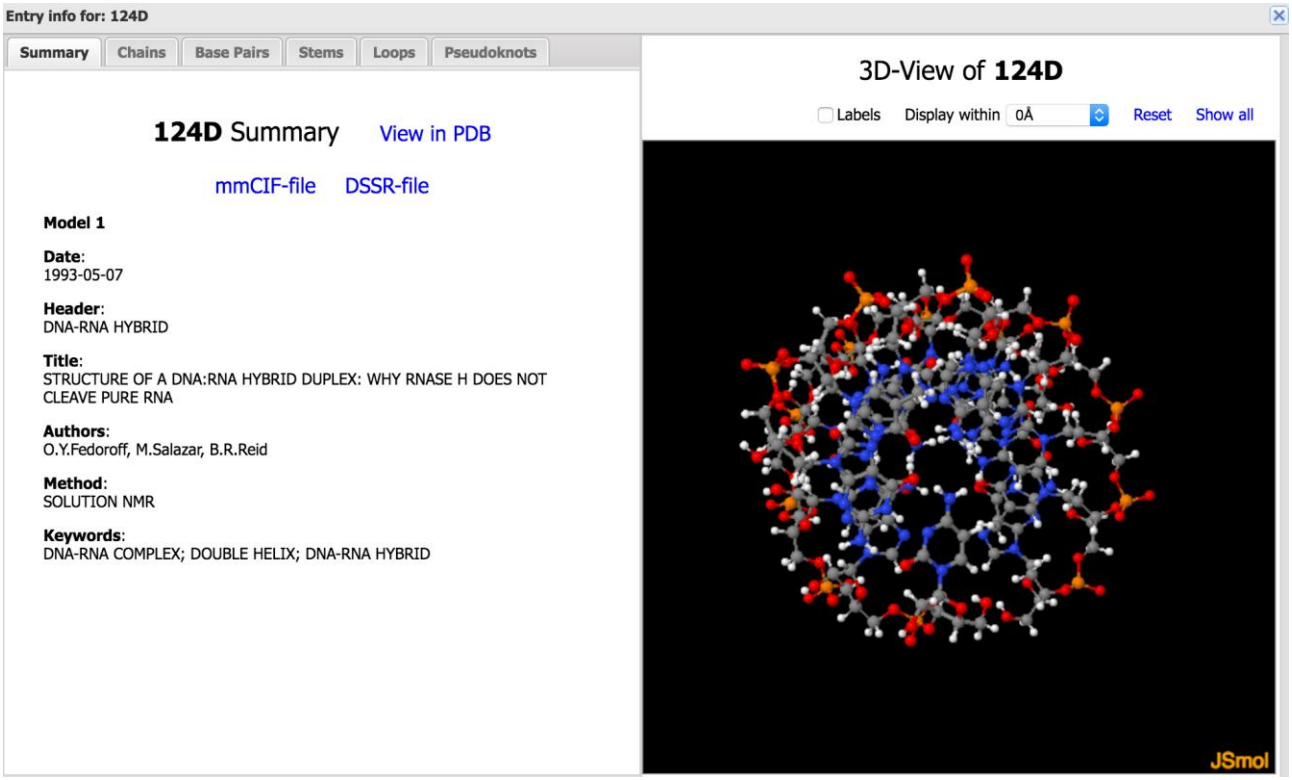


Рисунок №7

В окне со структурой присутствует множество разделов.

Summary. В данном разделе основная информация про структуру: автор, даты загрузки на сайт PDB, метод, которым были получены результаты.

Chains. Раздел, в котором указаны цепи в различных нотациях.

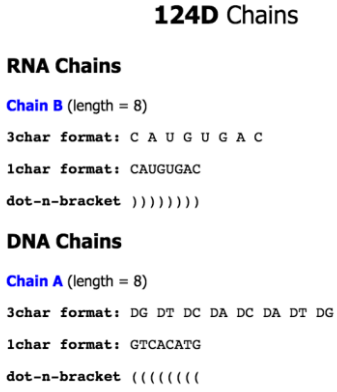


Рисунок №8

Base Pairs. Раздел с информацией о спаренных основаниях.

124D Base Pairs Help

No	Nucl1	Pair	Nucl2	Class by Saenger	Class by Leontis	Class by DSSR
1	A_1_DG	G-C	C_16_B	19-XIX	cWW	cW-W
2	A_2_DT	T-A	A_15_B	20-XX	cWW	cW-W
3	A_3_DC	C-G	G_14_B	19-XIX	cWW	cW-W
4	A_4_DA	A-G	G_12_B	n/a	cWW	cW-W
5	A_4_DA	A-U	U_13_B	20-XX	cWW	cW-W
6	A_5_DC	C-G	G_12_B	19-XIX	cWW	cW-W
7	A_6_DA	A-U	U_11_B	20-XX	cWW	cW-W
8	A_7_DT	T-A	A_10_B	20-XX	cWW	cW-W
9	A_8_DG	G-C	C_9_B	19-XIX	cWW	cW-W

Рисунок №9

Стемы. Раздел с информацией о стемах в структуре.

124D Stems Help

No	Length	#Wobble	Left	Right	Residues	
1	8	0	DG, DT, DC, DA, DC, DA, DT, DG	C, A, U, G, U, G, A, C	1-8:A, 9-16:B	view

Рисунок №10

Петли. (для структуры с PDB ID 1A1T_1)

1A1T Loops Help

Hairpins

No	Length	Type	Stem	Threads	Wings
1	4	classical	201-208:B, 213-220:B	209-212:B	

Description: 4 view

Рисунок №11

Псевдоузлы. (для структуры 1A60_1)

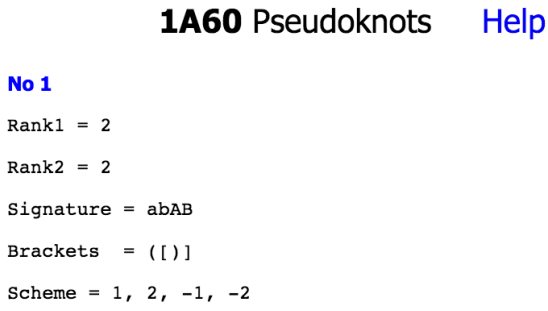


Рисунок №12

Раздел статистики

Данный раздел позволяет пользователям просматривать статистику, относящуюся к структурным элементам. Запрос может проходить в четырех режимах:

1. для всей базы данных
2. для выбранного набора структур
3. для неизбыточного PDB множества

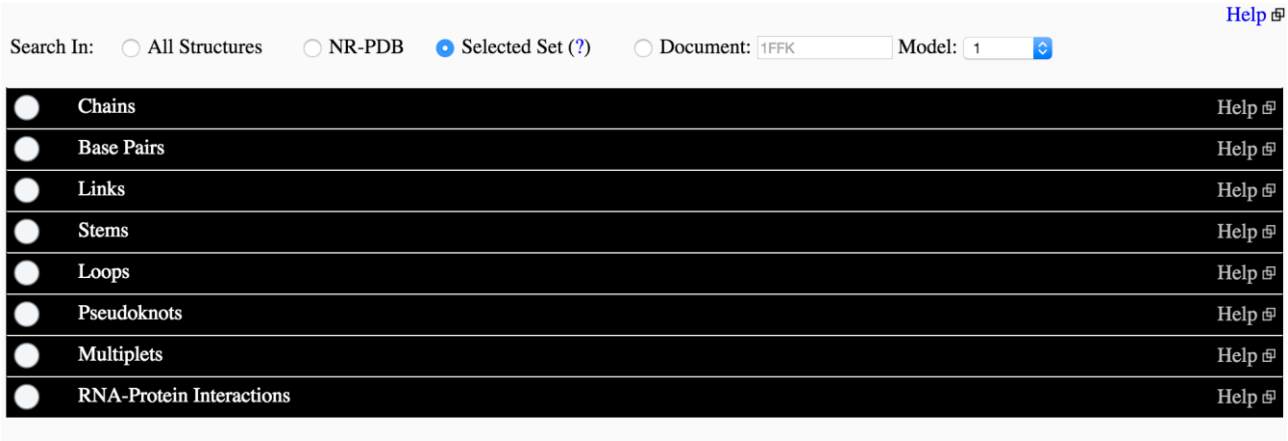
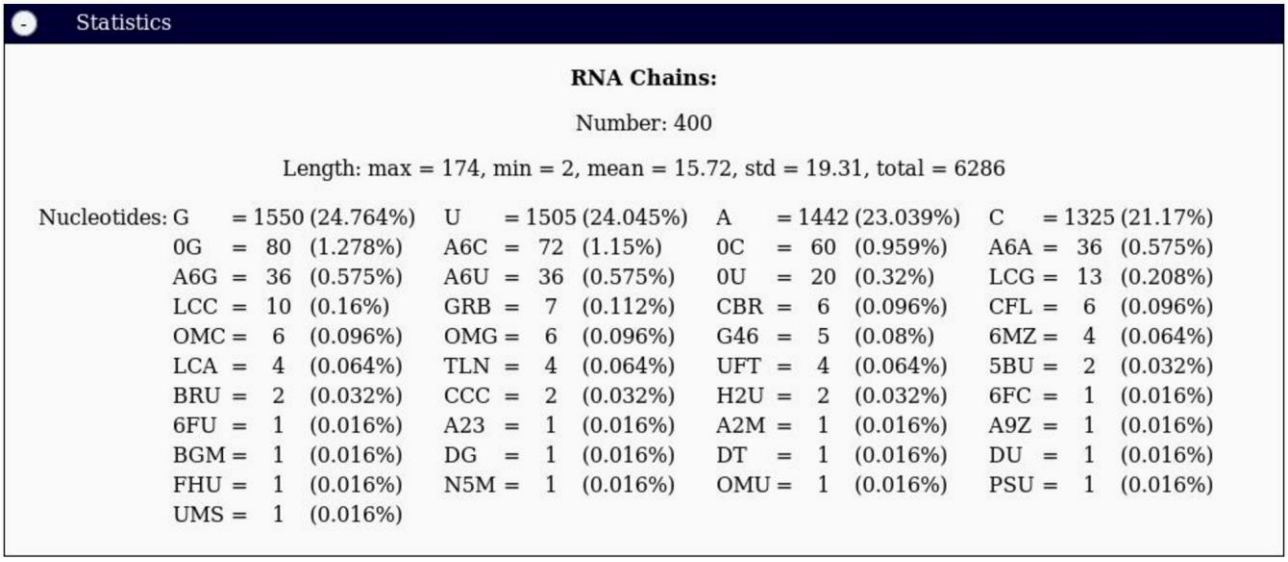


Рисунок №13

4. для выбранного PDB документа

После обработки запроса пользователя можно использовать различные фильтры для обработки результатов. Также есть возможность получить таблицу, содержащую полный список структурных элементов, которые соответствуют определенным условиям.

Пример результатов статистики по цепям:


Statistics

RNA Chains:

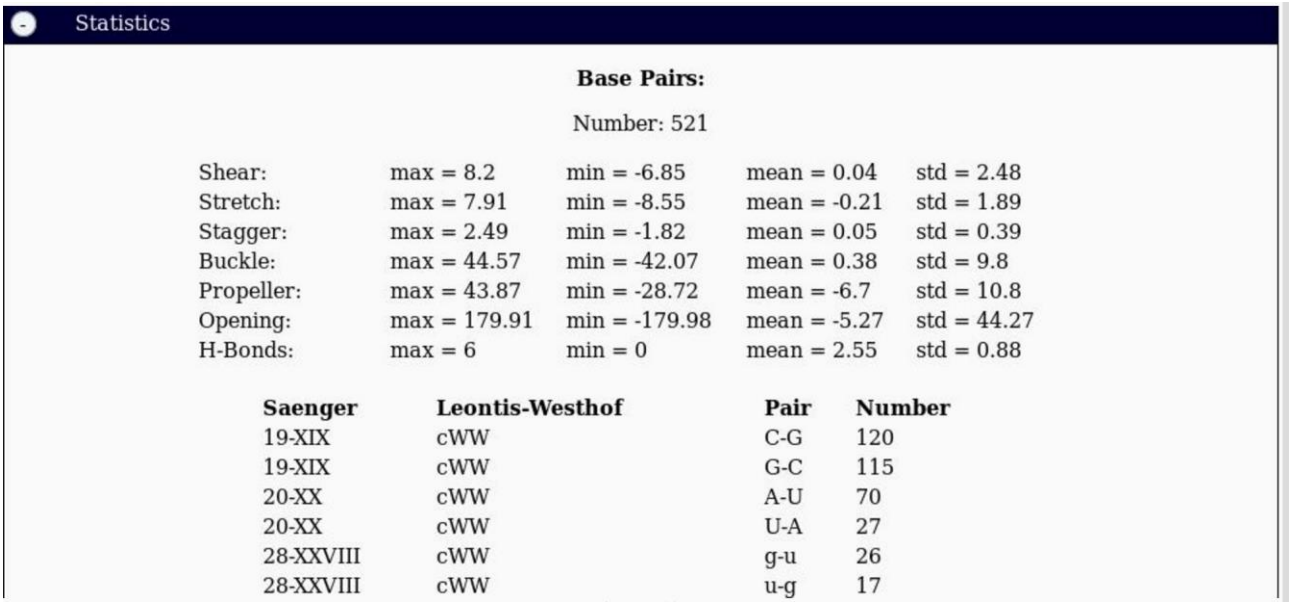
Number: 400

Length: max = 174, min = 2, mean = 15.72, std = 19.31, total = 6286

Nucleotides:	G	= 1550	(24.764%)	U	= 1505	(24.045%)	A	= 1442	(23.039%)	C	= 1325 (21.17%)
	0G	= 80	(1.278%)	A6C	= 72	(1.15%)	0C	= 60	(0.959%)	A6A	= 36 (0.575%)
	A6G	= 36	(0.575%)	A6U	= 36	(0.575%)	0U	= 20	(0.32%)	LCG	= 13 (0.208%)
	LCC	= 10	(0.16%)	GRB	= 7	(0.112%)	CBR	= 6	(0.096%)	CFL	= 6 (0.096%)
	OMC	= 6	(0.096%)	OMG	= 6	(0.096%)	G46	= 5	(0.08%)	6MZ	= 4 (0.064%)
	LCA	= 4	(0.064%)	TLN	= 4	(0.064%)	UFT	= 4	(0.064%)	5BU	= 2 (0.032%)
	BRU	= 2	(0.032%)	CCC	= 2	(0.032%)	H2U	= 2	(0.032%)	6FC	= 1 (0.016%)
	6FU	= 1	(0.016%)	A23	= 1	(0.016%)	A2M	= 1	(0.016%)	A9Z	= 1 (0.016%)
	BGM	= 1	(0.016%)	DG	= 1	(0.016%)	DT	= 1	(0.016%)	DU	= 1 (0.016%)
	FHU	= 1	(0.016%)	N5M	= 1	(0.016%)	OMU	= 1	(0.016%)	PSU	= 1 (0.016%)
	UMS	= 1	(0.016%)								

Рисунок №14

По спаренным основаниям:


Statistics

Base Pairs:

Number: 521

Shear:	max = 8.2	min = -6.85	mean = 0.04	std = 2.48
Stretch:	max = 7.91	min = -8.55	mean = -0.21	std = 1.89
Stagger:	max = 2.49	min = -1.82	mean = 0.05	std = 0.39
Buckle:	max = 44.57	min = -42.07	mean = 0.38	std = 9.8
Propeller:	max = 43.87	min = -28.72	mean = -6.7	std = 10.8
Opening:	max = 179.91	min = -179.98	mean = -5.27	std = 44.27
H-Bonds:	max = 6	min = 0	mean = 2.55	std = 0.88

Saenger	**Leontis-Westhof**	**Pair**	**Number**
19-XIX	cWW	C-G	120
19-XIX	cWW	G-C	115
20-XX	cWW	A-U	70
20-XX	cWW	U-A	27
28-XXVIII	cWW	g-u	26
28-XXVIII	cWW	u-g	17

Рисунок №15

По линкам:

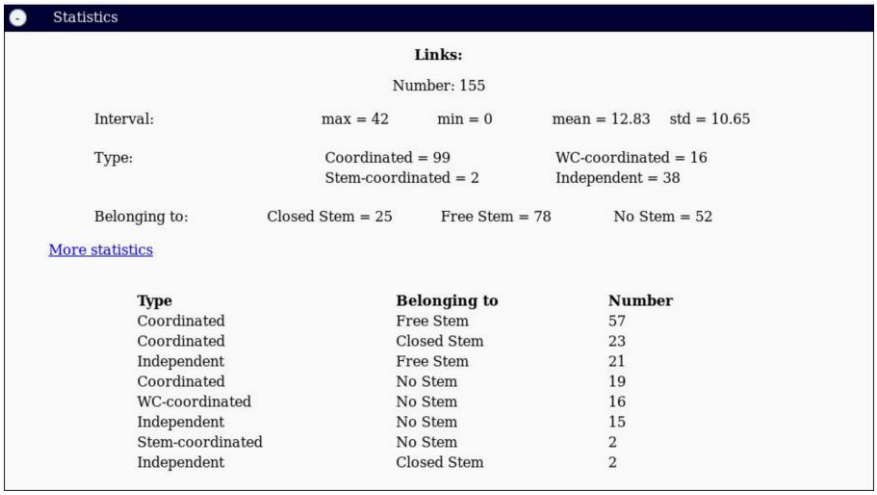


Рисунок №16

По стемам:

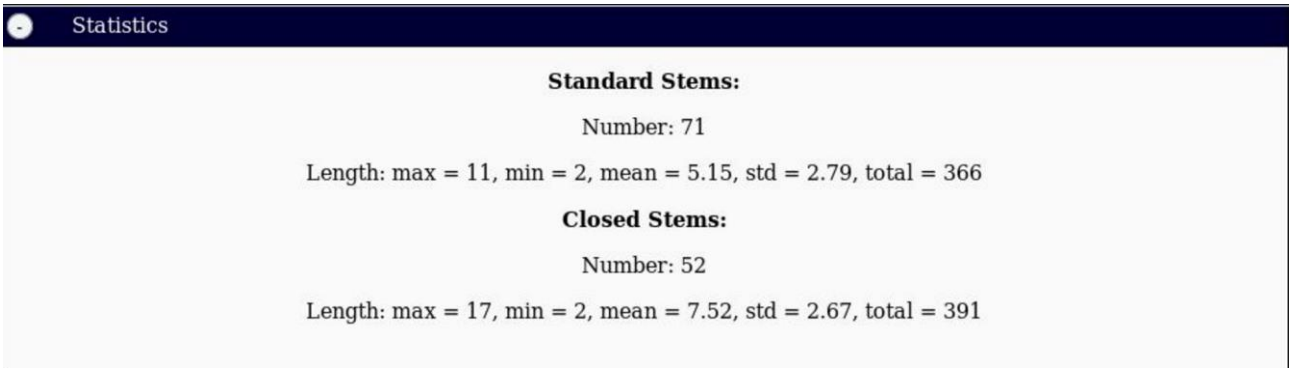


Рисунок №17

По петлям(«шпильки»):

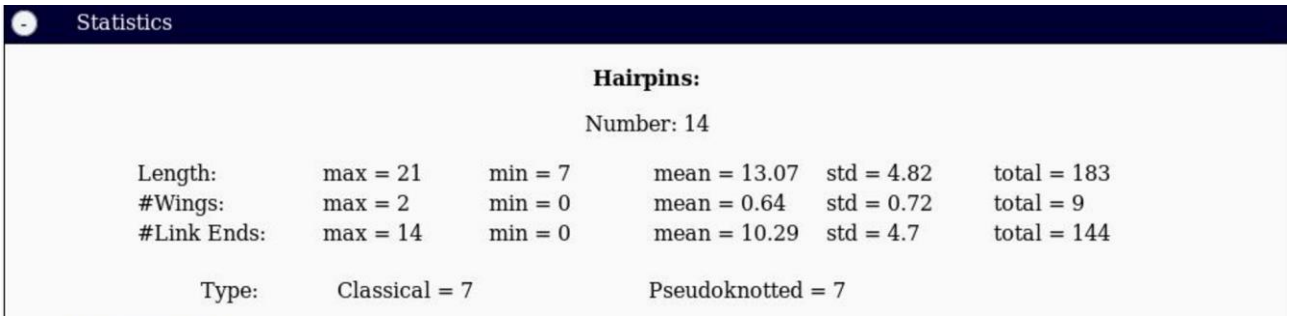


Рисунок №18

По псевдоузлам:

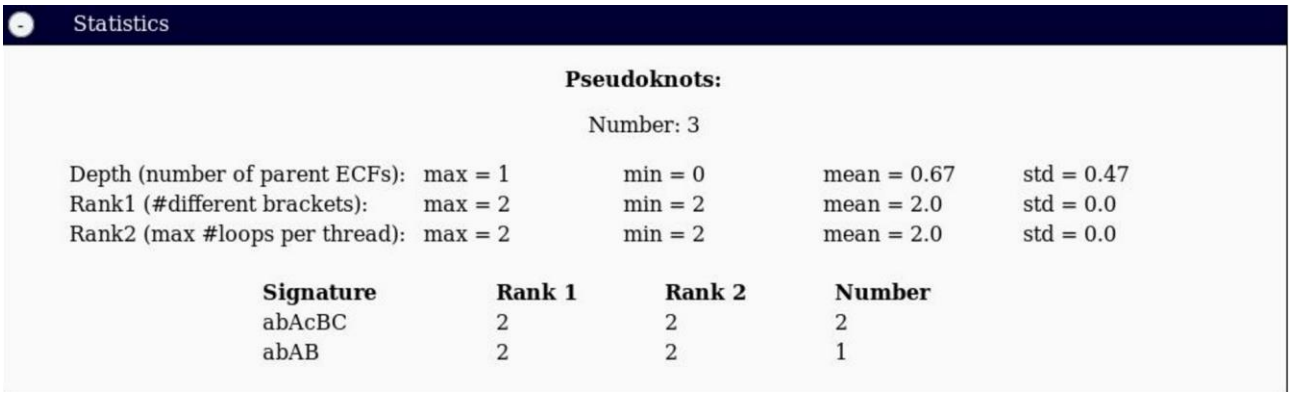


Рисунок №19

Также есть другие варианты фильтрации, которые могут быть изучены на официальном сайте.

Фреймворк

Предсказание перспективности любого наукоемкого стартапа это больше искусство нежели точный расчет, как это бывает в более классических бизнесах. Однако и тут используются классические сигналы для определения потенциальной перспективности. Основные из них:

- Рынок
- Наличие новизны в идее
- Профессионализм команды

Основные проблемы и задачи, с которыми сталкиваются стартапы, которые выходят на рынок достаточно большой, чтобы привлечь конкуренцию со стороны других стартапов и компаний, можно сформулировать следующим образом. Данные пункты являются модернизацией в сторону наукоемкости пунктов, которые выделяли в своей

работе «Predictors of success in new technology based ventures» Roure и Keeley[24]. В описании пункта будем показывать, как справляется с ними проект NR DB (если данный пункт применим к NR DB):

- *Определение актуальной проблемы на рынке (в научной области)*. Данный пункт выполнен, как демонстрировалось в анализе рынка.
- *Наличие необходимых средств*. В наукоемких стартапах основным ресурсом является людской. Профессиональная команда, занимающаяся данным проектом, является показателем того, что данная проблема решена.
- *Привлечение дополнительных ключевых сотрудников и достижение быстрого технического прогресса в новом продукте*
- *Определение ключевых клиентов и поставщиков.* Данный пункт выполнен в пункте, где показывалась актуальность проблемы.
- *Потенциальное получение дополнительного финансирования для увеличения штата и использования более передовых технологий*
- *Своевременная поправка бизнес стратегии на изменения рынка и от отзывов клиентво*
- *Увеличения продуктовой линии и увеличение компании.* Интеграция возможности сбора статистики с NR подмножеством продемонстрировано на сайте URSDB
- *Увеличенное время планирования*. Данное утверждение подтверждается в эмпирической работе «A profile of new venture success and failure in an emerging industry» Duchesneau и Gartner[25].

То, на сколько хорошо стартапы преодолевают данные проблемы и решают поставленные задачи и является хорошим фреймворком для определения его перспективности.

Актуальность проблемы и динамика рынка для NR базы данных была описана в начальных главах. В сфере статистических исследований в данной области существует спрос на данную базу данных по причине того, что избыточность данных делает эти исследования значительно не точными.

Соответственно, чтобы оценить потенциальную перспективность, нужно посмотреть, как текущая модель проекта вписывается в критерии фреймворка, который был описан выше.

---

[24] Roure J.B., Keeley R.H. Predictors of success in new technology based ventures // Journal of Business Venturing Vol. 5, Issue 4, 1990. PP. 201-220.

[25] Duchesneau D.A., Gartner W.B. A profile of new venture success and failure in an emerging industry // Journal of Business Venturing. Vol. 5, Issue 5, 1990. PP. 297-312.

Рынок & Научная актуальность

Выполнение данных критериев достигается за счет следующих факторов:

1. Рост рынка, который показан в основной части. Рост как с точки зрения исследовательской, так и венчурной.
2. Как было показано выше, с появлением огромного количества биологических данных необходимость в средствах анализа этих данных растет с огромной скоростью. На фоне улучшения механизмов определения 3D и вторичной структуры РНК и других биомолекул возможности ученых в изучении этих структур возросли. Однако до сих пор не существует удобной и надежной платформы, которая позволяет изучать РНК структуры в неизбыточном множестве. Отсюда и появилась очевидная необходимость в создании соответствующего инструмента.

Как уже было показано во многих стартапах, которые создавались на основе похожий идей, которые пытались реализовывать другие, новизна идеи является важным фактором, но далеко не решающим. Так способность команды реализовать MVP продукт в относительно сжатые сроки показывает степень профессионализма команды, а также потенциальную трудоемкость в полноценной реализации идеи. Описание уже функционирующего сайта URSDB показывает способность команды создавать продукт, что является еще одним важным критерием в определении перспективности проекта.

Заключение

С ростом рынка биотехнологических стартапов, а также с увеличением количества биологических данных, которые появляются каждый день, актуальность получения доступа к инструментам по работе с данными, обладающими свойством неизбыточности, растет.

Данные предпосылки привели к необходимости создания NR базы данных РНК структур. РНК структуры были выбраны в качестве целевых по причине того, что в отличие от белков и ДНК данная биологическая структура является «обделенной». Однако, это является неприемлемым, поскольку РНК играет ключевую роль во многих новейших исследованиях. Отличным примером является стартап Cofactor Genomics, который занимается тем, что использует РНК для диагностирования различных заболеваний.

В ходе данной исследовательской работы были показаны основные теоретические и практические аспекты, связанные с созданием NR базы данных. Их кратко можно обозначить следующим образом:

- Получение данных для кластеризации из открытых источников различными программными способами. Нами использовался крупнейший сайт в этой области - PDB.
- Создание вспомогательных инструментов
- Выработка критериев для кластеризации
- Составление эффективного алгоритма кластеризации
- Создание механизма для удобного взаимодействия с результатами работы

После описания того, как создавалась NR база данных были описаны критерии, которые составляют фреймворк для определения перспективности стартапа на данном рынке. Анализ проекта относительно составленных пунктов позволил сделать предположение о перспективности данного наукоемкого стартапа.

Перед будущими исследователями стоят следующие потенциальные проблемы:

- Создание экономической модели для численной оценки моделей монетизации
- Использование эконометрических моделей для подбора критериев для фреймворка по определению перспективности

- Расширение фреймворка на весь рынок стартапов в биоинофрмационных технологиях

Список литературы

1. 23andMe. 2016. URL: https://www.23andme.com/en-int/ (дата обращения: 18.04.2016).
2. Acs, Z.J., Audretsch, D.B., 1988. Innovation in large and small firms: an empirical analysis // American Economic Review. 1988. Vol. 78. № 4. PP. 678-690.
3. Benchling. 2015. URL: https://benchling.com/ (дата обращения:24.04.2016).
4. Bioinformatics Market Analysis By Product (Sequence Analysis, Manipulation, Alignment, Structural & Functional Analysis Platforms, Data Management, Sequencing, Data Analysis Service, Generalized, Specialized Biocontent), By Application (Genomics, Molecular Phylogenetics, Metabolomics, Proteomics, Transcriptomics, Chemoinformatics & Drug Designing) And Segment Forecasts To 2020: report summary // Grand View Research, Inc. 2015. August. URL: http://www.grandviewresearch.com/industry-analysis/bioinformatics-industry (дата обращения: 18.04.2016).
5. Chan, L. K., Lakonishok J., Sougiannis T. The stock market valuation of research and development expenditures // The Journal of Finance. 2001. Vol. 56. № 6. PP. 2431-2456.
6. Cock P.J.A., et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics // Bioinformatics. 2009. Vol. 25. № 11. PP. 1422-1423. doi:10.1093/bioinformatics/btp163
7. Cofactor Genomics. 2013. URL: https://cofactorgenomics.com/ (дата обращения:24.04.2016).
8. de Hoon M.J.L., et al. Open source clustering software // Bioinformatics. 2004. Vol. 20. № 9. PP. 1453-1454. doi:10.1093/bioinformatics/bth078
9. Dibner M.D., Trull M., Howell M. US venture capital for biotechnology // Nature Biotechnology. 2003. Vol. 21. PP. 613 – 617.
10. DNAnexus. 2015. URL: https://www.dnanexus.com/ (дата обращения: 25.04.2016).
11. Duchesneau D.A., Gartner W.B. A profile of new venture success and failure in an emerging industry // Journal of Business Venturing. Vol. 5, Issue 5, 1990. PP. 297-312.

12. Ehrenberg R. Creating competitive advantage through data // IA Ventures' blog. 2011. URL: http://fortune.com/2011/07/21/creating-competitive-advantage-through-data/ (дата обращения: 10.05.2016).
13. Gershon D. Bioinformatics in a post-genomics age // Nature. 25.09.1997. Vol. 389. PP. 417-418.
14. Good B.M., Su A. J. Crowdsourcing for bioinformatics // Department of Molecular and Experimental Medicine, The Scripps Research Institute. La Jolla, 2013. doi: 10.1093/bioinformatics/btt333
15. Human Genome Project Information Archive: 1990-2003 // U.S. DOE Human Genome Project. Дата обновления: 09.05.2016. URL: http://web.ornl.gov/sci/techresources/Human_Genome/ (дата обращения: 10.05.2016).
16. Kawrykow A. et al. Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment // PLoS ONE. 2012. Vol. 7. № 3: e31362. doi:10.1371/journal.pone.0031362
17. Leontis N.B. Zibrel C.L. Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking // Nucleic Acids and Molecular Biology. New York, 2012. Vol. 27. PP. 281-298.
18. Leontis N.B., Westhof E. Geometric nomenclature and classification of RNA base pairs // RNA. 2001. Vol. 7. № 4. PP. 499–512.
19. Leslie A.J., Philippe C.W. The determinants of venture capital funding: evidence across countries **//** Journal of Corporate Finance. 2000. Vol. 6. Issue 3. PP. 241–289.
20. Margaret Oakley Dayhoff 1925–1983 // Bulletin of Mathematical Biology. 1984. Vol. 46. Issue 4. PP. 467-472.
21. Methods for Monetizing Your Data [Электронный ресурс]: URL: http://www.gartner.com/webinar/3098518 (дата обращения: 05.05.2016).
22. Notable Labs. 2016. URL: https://notablelabs.com/ (дата обращения:24.04.2016).
23. Reed J. Trends in Commercial Bioinformatics // Oscar Gruss Biotechnology review. New York, 2000. №13.
24. Reinganum, J.R. The timing of innovation: research, development, and diffusion // Handbook of Industrial Organization. 1989. Vol. 1. PP. 849-908.
25. Roure J.B., Keeley R.H. Predictors of success in new technology based ventures // Journal of Business Venturing Vol. 5, Issue 4, 1990. PP. 201-220.

26. Sanger F. The free amino groups of insulin // Biochemical Journal. 1945. Vol. 39. № 5. PP. 507-515. doi:10.1042/bj0390507
27. Science Exchange. 2016. URL: https://www.scienceexchange.com/ (дата обращения:24.04.2016).
28. Soylent. 2016. URL: https://www.soylent.com/ (дата обращения: 24.04.2016).
29. The Protein Data Bank: [Электронный ресурс] // Research Collaboratory for Structural Bioinformatics. 2003. URL: http://www.rcsb.org/pdb/home/home.do (дата обращения: 15.04.2016).
30. Transcriptic. 2015. URL: https://www.transcriptic.com/ (дата обращения: 24.04.2016).

Приложения

Приложение № 1

```
import os, random, time, sys, csv
from openpyxl import Workbook
from openpyxl.cell import get_column_letter
from Bio import pairwise2
from random import shuffle
from Bio.Seq import Seq
from Bio.Alphabet import IUPAC

args = sys.argv


aligment_type = args[1]
input_file = args[2]
output_file = args[3]


def read_seq():
    fname = os.path.join('', input_file)
    data = []
    with open(fname, 'r') as f:
        for line in f:
            vs = line.strip().split(';')
            nucleo = str(vs[4]).strip().split(',')
            res = (str(vs[0]), nucleo)
            data.append(res)
    del(data[0])
    return data

def find_score(seq1, seq2, alignment):
    l1, l2 = len("".join(seq1)),len("".join(seq2))
    l = min(l1,l2)
    if alignment == "local":
```

```
        _,_, score, _, _ = (pairwise2.align.localxs("".join(seq1), "".join(seq2),-.5, -
.1))[0]
    elif alignment == "global":
        _,_, score, _, _ = (pairwise2.align.globalxs("".join(seq1), "".join(seq2),-.5, -
.1))[0]
    else:
        raise ValueError("Incorrect type of alignment")
    return score/l

def makeCluster(clusters, elem):
        clusters.append([elem])

def findCluster(target, clusters):
        for cluster in clusters:
                seq = random.choice(cluster)
                score = find_score(seq[1], target[1], aligment_type)
                if score > 0.8:
                        cluster.append(target)
                        return
        makeCluster(clusters, target)
        return


def cluster(seqs, clusters):
        shuffle(seqs)
        if len(seqs) < 1:
                raise ValueError('Not enough sequences')
        makeCluster(clusters,seqs[0])
        if len(seqs) == 0:
                return clusters

        for seq in seqs[1:]:
                findCluster(seq, clusters)

        return


def write_clusters(data):
        res = []
```

```
        for indx in range(0,len(data)):
                ids = []
                for cluster in data[indx]:
                        ids.append(cluster[0])
                res.append([indx+1, ",".join(ids)])
        names = [["cluster_id", "sequences"]]
        res = names + res
        with open(output_file, "w") as f:
                writer = csv.writer(f, delimiter=';')
                writer.writerows(res)
        wb = Workbook()
        ws1 = wb.active
        ws1.title = "input"
        for row_index, row in enumerate(csv.reader(open(input_file, 'r'), delimiter=';')):
            for column_index, cell in enumerate(row):
                column_letter = get_column_letter((column_index + 1))
                ws1.cell('%s%s'%(column_letter, (row_index + 1))).value = cell

        ws2 = wb.create_sheet(title="output")
        for row_index, row in enumerate(csv.reader(open(output_file, 'r'), delimiter=';')):
            for column_index, cell in enumerate(row):
                column_letter = get_column_letter((column_index + 1))
                ws2.cell('%s%s'%(column_letter, (row_index + 1))).value = cell
        wb.save("results.xls")




clusters = []
seqs = read_seq()

cluster(seqs, clusters)
clusters = sorted(clusters, key=len, reverse=True)
write_clusters(clusters)
print("Success!")
```

Приложение № 2

```
import os, random, time, sys, csv
from openpyxl import Workbook
from openpyxl.cell import get_column_letter
from collections import defaultdict


args = sys.argv


input_file = args[1]
none_val = "none && none"

def read_seq(condition):
	if not condition:
		fname = os.path.join('', input_file)
		data = []
		with open(fname, 'r') as f:
			for line in f:
				vs = line.strip().split(';')
				molecule = str(vs[6]).strip().lower()
				if "&&" in molecule:
					fragments = str(vs[8]).strip().lower()
					if "&&" in fragments:
						if none_val not in fragments:
							#x--z && y--w
							tmp_mol = molecule.split('&&')
							tmp_frag = fragments.split('&&')
							res = []
							bind1 = "--"
							bind2 = "&&"
							for ind in range(0,len(tmp_mol)):
								tmp = tmp_mol[ind] + bind1 +
tmp_frag[ind]
								res.append(tmp)
							molecule = bind2.join(res)
				org = str(vs[7]).strip().lower()
				data.append((molecule, org))
		del(data[0])
	else:
```

```
        fname = os.path.join('', input_file)
        data = []
        with open(fname, 'r') as f:
            for line in f:
                vs = line.strip().split(';')
                pbd_id = str(vs[1]).strip()
                suffix = "_1"
                if pbd_id.endswith(suffix):
                    molecule = str(vs[6]).strip().lower()
                    if "&&" in molecule:
                        fragments = str(vs[8]).strip().lower()
                        if "&&" in fragments:
                            if none_val not in fragments:
                                #x--z && y--w
                                tmp_mol = molecule.split('&&')
                                tmp_frag = fragments.split('&&')
                                res = []
                                bind1 = " -- "
                                bind2 = " && "
                                for ind in range(0,len(tmp_mol)):
                                    tmp = tmp_mol[ind] + bind1 + tmp_frag[ind]
                                    res.append(tmp)
                                molecule = bind2.join(res)
                    org = str(vs[7]).strip().lower()
                    data.append((molecule, org))
        del(data[0])
    return data


def write_stats(mol,org, agr, fileName):
    names_mol = [["molecule", "count"]]
    names_org = [["organisme", "count"]]
    names_agr = [["organisme", "molecule", "count"]]
    newagr = []
    for t in agr:
        tmp = []
```

```
                spl = t[0].split("+++++")
                tmp = [spl[0], spl[1], t[1]]
                newagr.append(tmp)
        mol = names_mol + mol
        org = names_org + org
        newagr = names_agr + newagr
        wb = Workbook()
        ws1 = wb.active
        ws1.title = "input"
        for row_index, row in enumerate(csv.reader(open(input_file, 'r'), delimiter=';')):
            for column_index, cell in enumerate(row):
                column_letter = get_column_letter((column_index + 1))
                ws1.cell('%s%s'%(column_letter, (row_index + 1))).value = cell

        ws2 = wb.create_sheet(title="molecules")
        for row_index, row in enumerate(mol):
            for column_index, cell in enumerate(row):
                column_letter = get_column_letter((column_index + 1))
                ws2.cell('%s%s'%(column_letter, (row_index + 1))).value = cell

        ws3 = wb.create_sheet(title="organismes")
        for row_index, row in enumerate(org):
            for column_index, cell in enumerate(row):
                column_letter = get_column_letter((column_index + 1))
                ws3.cell('%s%s'%(column_letter, (row_index + 1))).value = cell

        ws4 = wb.create_sheet(title="organisme&molecule")
        for row_index, row in enumerate(newagr):
            for column_index, cell in enumerate(row):
                column_letter = get_column_letter((column_index + 1))
                ws4.cell('%s%s'%(column_letter, (row_index + 1))).value = cell
        wb.save(fileName)

def freqMol(data):
        d = {}
        for term in data:
                if term[0] in d:
                        d[term[0]] += 1
```

```
                else:
                        d[term[0]] = 1
        return d

def freqOrg(data):
        d = {}
        for term in data:
                if "&&" in term[1]:
                        l = (term[1].split("&&")[0]).strip()
                        if l in d:
                                d[l] +=1
                        else:
                                d[l] = 1
                else:
                        if term[1] in d:
                                d[term[1]] += 1
                        else:
                                d[term[1]] = 1
        return d

def freqBoth(data):
        d = {}
        for term in data:
                if "&&" in term[1]:
                        l = (term[1].split("&&")[0]).strip()
                        k = l + "+++++" + term[0]
                        if k in d:
                                d[k] +=1
                        else:
                                d[k] = 1
                else:
                        k = term[1] + "+++++" + term[0]
                        if k in d:
                                d[k] += 1
                        else:
                                d[k] = 1
        return d

def toList(dct):
```

```
        ls = []
        for key, value in dct.items():
                tmp = [key, value]
                ls.append(tmp)
        ls.sort(key= lambda x:x[1], reverse=True)
        return ls

def slow_print(dta):
        for el in dta:
                print(el)
                time.sleep(2)

dta = read_seq(False)
resultName = "stats.v2.xls"
write_stats(toList(freqMol(dta)), toList(freqOrg(dta)), toList(freqBoth(dta)),
resultName)

condition_data = read_seq(True)
resultOnlyFirst = "statsPBD_1.v2.xls"
write_stats(toList(freqMol(condition_data)), toList(freqOrg(condition_data)),
toList(freqBoth(condition_data)), resultOnlyFirst)

print("Success!")
```