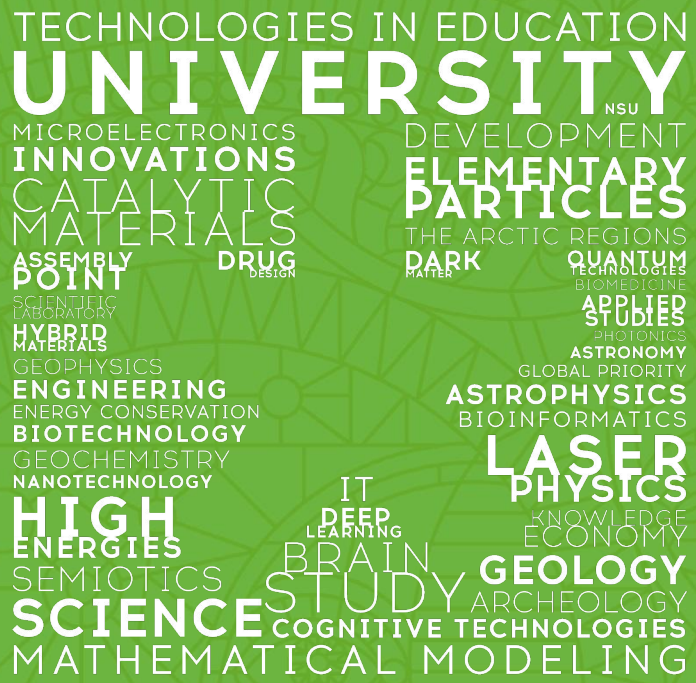


Novosibirsk State University
*THE REAL SCIENCE

Отчет

о прохождении <u>учебной практики,</u>
<u>научно-исследовательской работы</u>

Тема задания:
Изучение информационных источников и баз данных, а также методов и программных средств интеграции и анализа данных по белок-белковым взаимодействиям, экспрессии генов, биологическим онтологиям.

Студент: Филонов Сергей Вячеславович

Руководитель: Игнатьева Елена Васильевна
Консультант: Подколодный Николай Леонтьевич

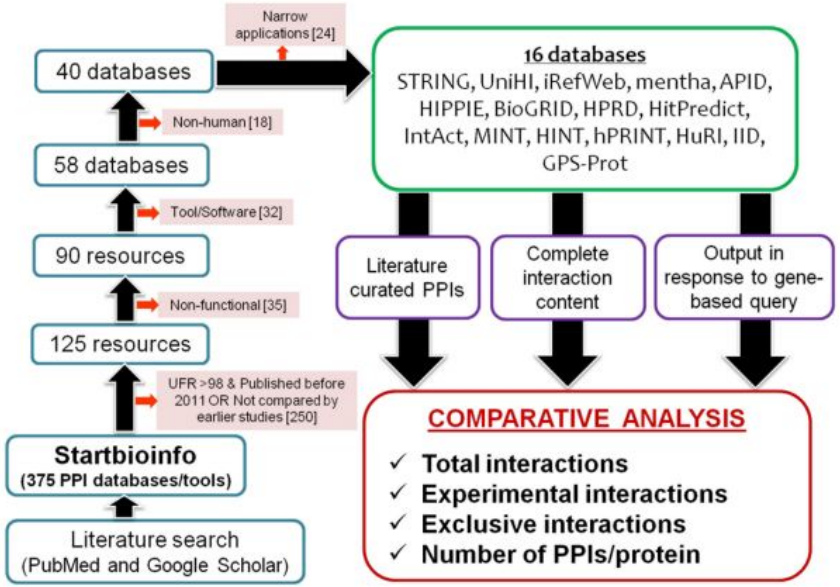
*Цели

- Ознакомление с основными направлениями исследований Отдела системной биологии ИЦиГ СО РАН, включая анализ сетей белок- белковых взаимодействий, анализ динамических биологических сетей, анализ суточной динамики биологических процессов, изучение объекта исследования (сети белок белковых взаимодействий).
- Составление аналитического обзора по информационным источникам и базам данных по белок-белковым взаимодействиям и экспрессионным данным, форматам представления данных и методам доступа к ним, методам анализа такого типа данных и программному обеспечению.
- Выбор и обоснование методик и средств решения поставленной задачи. Освоение программного обеспечения и библиотек для доступа к биоинформационным данным по белок белковым взаимодействиям, экспрессионным данным, а также методов и программного обеспечения по анализу такого типа данных .

*Задачи

- Выполнить аналитический перевод:
  - Bajpai AK, Davuluri S, Tiwary K, Narayanan S, Oguru S, Basavaraju K, Dayalan D, Thirumurugan K, Acharya KK. Systematic comparison of the protein-protein interaction databases from a user's perspective. J Biomed Inform. 2020 Mar;103:103380. doi: 10.1016/j.jbi.2020.103380. Epub 2020 Jan 28. PMID: 32001390.
  - Zahiri J, Emamjomeh A, Bagheri S, Ivazeh A, Mahdevar G, Sepasi Tehrani H, Mirzaie M, Fakheri BA, Mohammad-Noori M. Protein complex prediction: A survey. Genomics. 2020 Jan;112(1):174-183. doi: 10.1016/j.ygeno.2019.01.011. Epub 2019 Jan 17. PMID: 30660789.
- Скачать данные по мышам из баз данных STRING, IntAct, mentha, APID, MINT, IID.
- Скачать базу CORUM по белковым комплексам.
- Скачать в базе IntAct базу по белковым комплексам.
- Установить Cytoscape, изучить основные функции.
- Установить и научиться применять плагины для Cytoscape по прогнозированию белковых комплексов.

*Отбор баз данных



- Были исключены из обзора базы данных состоящие из данных полученных программно, с использованием систем интеллектуального анализа текста, не содержащие информацию о белок-белковых взаимодействиях у человека, имеющие высокую специфичность применения.

*Базы данных отобранные для детального сравнения

Database	URL	Version used	Update frequency	Primary/ Secondary	Data type (E/P)	Complete data download	Organism (number of organisms)
STRING [20]	http://string-db.org	v10.5	~1–1.5y	Secondary	E & P	Yes	Multiple (> 10)
UniHI [21]	http://www.unihi.org	v7.1	NF	Secondary	E & P	Yes[2]	Human[3]
mentha [22]	http://mentha.uniroma2.it/	–	1d*	Secondary	E	Yes	Multiple (> 10)
hPRINT [23]	http://print-db.org/hprint-web/	v10.1	NF	Secondary	E & P	Yes	Human
APID [24]	http://apid.dep.usal.es/	vJune 2017	~1y	Secondary	E	Yes	Multiple (> 10)
HIPPIE [25]	http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/	v2.1	~0.5-1y	Secondary[1]	E	Yes	Human
GPS-Prot [26]	http://www.gpsprot.org	v3.0.5	~1-2y	Secondary[1]	E	No	Human & HIV
BioGRID [27]	http://thebiogrid.org	v3.4.154	1 m	Primary	E	Yes	Multiple (> 10)
HPRD [28]	http://www.hprd.org	v9	NF	Primary	E	Yes[2]	Human
HitPredict [29]	http://hintdb.hgc.jp/htp/	v4	~1y	Secondary	E	Yes	Multiple (> 10)
IntAct [30]	http://www.ebi.ac.uk/intact	v4.2.10	~1y	Primary	E	Yes	Multiple (> 10)
MINT [31]	http://mint.bio.uniroma2.it	–	NF	Primary	E	Yes	Multiple (> 10)
HINT [32]	http://hint.yulab.org/	v4	1d*	Secondary	E	Yes	Multiple (> 10)
IID [33]	http://ophid.utoronto.ca/iid	v2017-04	NF	Secondary	E & P	Yes[2]	Multiple (< 10)
HuRI	http://interactome.baderlab.org/	–	NF	Primary	E	Yes[2]	Human
iRefWeb [34]	http://wodaklab.org/iRefWeb/	v13.0	~1m-1y	Secondary	E & P	Yes	Multiple (> 10)

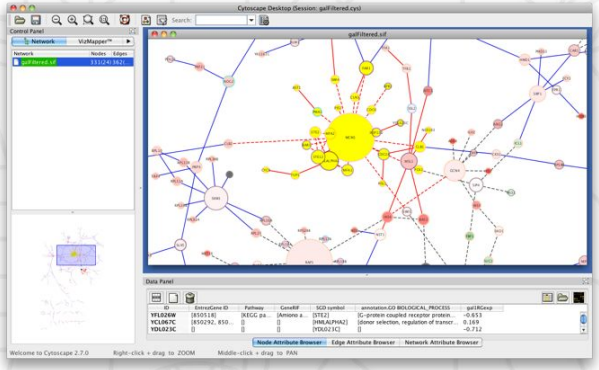
N*

Методы сравнения

- Сравнение оставшихся баз данных производилось по наличию дополнительных функций запросов, охвату “золотого стандарта”, количеству взаимодействий и белков содержащихся в базе данных, количеству эксклюзивных взаимодействий и белков для каждой базы данных.

* Визуализация и обработка данных

Для визуализации и обработки данных была выбрана программа Cytoscape.



- Cytoscape - биоинформатическая платформа с открытым исходным кодом, предназначенная для визуализации сетей молекулярных взаимодействий и биологических путей с возможностью использования дополнительных данных, таких как функциональная аннотация, информация об уровне экспрессии генов и прочих. Несмотря на то, что исходно Cytoscape был разработан для биологических исследований, сейчас он широко используется для решения различных задач по анализу сетей и их визуализации.

Белковые комплексы

- Некоторые подмножества взаимодействующих белков в сетях белок-белковых взаимодействий образуют комплексы. Белковые комплексы являются одной из важнейших функциональных единиц для протекания биологических процессов в клетке. Эти белковые комплексы принимают участие в различных биологических процессах в том числе: регуляции клеточного цикла, дифференцировке, сворачивание белков, трансляции, транскрипции, посттрансляционной модификации, экспрессии генов, ингибирование ферментов и антитело-антигенных взаимодействиях. Существует ряд вычислительных методов для предсказания белковых комплексов по данным о бинарных взаимодействиях белков. Вычислительные методы прогнозирования белковых комплексов можно разделить на три основные категории: сетевые, биологически-контекстно-зависимые и специализированные.

*Отбор баз данных для поиска белковых комплексов

- Для поиска белковых комплексов из вышеперечисленных баз данных были отобраны содержащие информацию о белок белковых взаимодействиях у мышей(STRING, IntAct, mentha, APID, MINT, IID).

* Плагины и реализованные в них алгоритмы для предсказания белковых комплексов

- ClusterONE
  - Выбор белка с самой высокой степенью в качестве начальной позиции
  - Отбор группы белков с самой высокой “сплоченностью” содердащей выбронный белок
  - Выбирает следующий белок, рассматривая все белки, которые не были отнесены к найденным досих пор белковым комплексам.
- MCODE
  - Взвешивание узлов
  - Прогнозирование комплексов
  - Добавление/ удаление белков в/из обнаруженных комплексов в соответствии с критерием связности.

* Плагины и реализованные в них алгоритмы для предсказания белковых комплексов

- CytoClaster
  - HC-PIN
    - Все вершины в сети PPI рассматриваются как одноэлементные кластеры
    - Вычисляет значение кластеризации каждого ребра и помещает все ребра в очередь в порядке невозрастания в соответствии с их значениями кластеризации. Чем выше значение кластеризации ребра, тем больше вероятность, что две его вершины будут в одном модуле.
    - Постепенно добавляя ребра из очереди к кластерам, алгоритм собирает все одноэлементные кластеры в λ –модули.
    - λ -модули могут быть выведены, когда количество его белков не меньше порогового значения s.

Algorithm HC-PIN
input: Undirected graph $G = (V, E)$, parameter λ and s;
output: identified modules;
1. **for** each vertex $v_i \in V$ **do**
   $C_i = \{\{v_i\}, \emptyset\}$
//each vertex v_i is initialized as a singleton cluster.
2. **for** each edge $(u, v) \in E$ **do**
   compute its clustering value;
3. sort all edges to queue S_q in non-increasing order in terms of their clustering values;
4. **while** $S_q \neq \emptyset$ **do**
   $\{(u, v) \leftarrow S_q$; //the first edge (u,v) in S_q is selected.
   if $L(u) = L(v)$ //u and v are in the same cluster
   then $i = L(u)$; $E_i = E_i \cup \{(u, v)\}$;
   else $i = L(u)$; $j = L(v)$;
   if $\frac{D_{in}(C_i)}{D_{out}(C_i)} \leq \lambda$ or $\frac{D_{in}(C_j)}{D_{out}(C_j)} \leq \lambda$ **then**
   $V_i = V_i \cup V_j$;
   $E_i = E_i \cup E_j \cup \{(u, v)\}$;
   $C_j = \{\emptyset, \emptyset\}$;
   $S_q = S_q - (u, v)$; //edge (u,v) is removed from $S_q\}$
5. **for** i=1 to $|V|$ **do**
   if $|V_i| \geq s$ **then** output C_i;
//output all the clusters consisting of at least s proteins.

* Плагины и реализованные в них алгоритмы для предсказания белковых комплексов

- OH-PIN
  - Вначале множество кластеров C_set пусто.
  - Для каждого ребра в сети взаимодействия белков создается его B_Cluster, и B_Cluster добавляется в C_set, если B_Cluster еще не включен в C_set, до тех пор, пока не будет включен каждый B_Cluster.
  - Затем OH-PIN объединяет все сильно перекрывающиеся пары кластеров в C_set исходя из заданного порогового значения перекрытия.
  - Собирает все кластеры в C_set в λ-модули, постепенно объединяя пары кластеров с максимальным коэффициентом кластеризации.
- IPCA
  - Вычисляет вес каждого ребра, подсчитывая общих соседей двух связанных узлов, и вычисляет вес каждого узла, суммируя веса его инцидентных ребер.
  - Вершина с наибольшим весом будет рассматриваться как начальный кластер
  - Расширяет кластер, рекурсивно добавляя вершины из его соседей с точки зрения приоритета узлов. Возможность добавления узла в кластер определяется двумя условиями: вероятностью его взаимодействия и кратчайшим путем между ним и узлами в кластере.

Algorithm IPCA
input: a graph $G = (V, E)$, parameters T_{in} and d;
output: identified complexes (clusters);

(** Weighting Vertex **)
1. compute the weight of each edge;
2. compute the weight of each vertex;
3. queue $S_q \leftarrow$ all vertices sorted in non-increasing order in terms of vertex weights;

(** Selecting Seed **)
4. **while** $S_q \neq \emptyset$ **do** $\{ v \leftarrow S_q;\ K = \{v\};$ call ExtendingCluster(K). $\}$

Subroutine ExtendingCluster(K) (** Extending Cluster **)

(** Extend-judgement **)
1. **if** there is a (K, T_{in}, d)-vertex
2. **then** let v be a (K, T_{in}, d)-vertex that has the highest priority;
   Call ExtendingCluster($K + v$);
3. **else** print the cluster K; $S_q = S_q - K$.

Разработка плагинов

- Cytoscape позволяет интегрировать в платформу собственные плагины. Был изучен ее API, разобран исходный код плагинов ClusterONE, MCODE. Получены навыки разработки плагинов для Cytoscape.

Заключение

- Изучены следующие статьи:
  - Bajpai AK, Davuluri S, Tiwary K, Narayanan S, Oguru S, Basavaraju K, Dayalan D, Thirumurugan K, Acharya KK. Systematic comparison of the protein-protein interaction databases from a user's perspective. J Biomed Inform. 2020 Mar;103:103380. doi: 10.1016/j.jbi.2020.103380. Epub 2020 Jan 28. PMID: 32001390.
  - Zahiri J, Emamjomeh A, Bagheri S, Ivazeh A, Mahdevar G, Sepasi Tehrani H, Mirzaie M, Fakheri BA, Mohammad-Noori M. Protein complex prediction: A survey. Genomics. 2020 Jan;112(1):174-183. doi: 10.1016/j.ygeno.2019.01.011. Epub 2019 Jan 17. PMID: 30660789.
- Проведен сравнительный анализ и отбор баз данных по белок-белковым взаимодействиям и белковым комплексам для создания интегрированной базы данных.
- Скачены и подготовлены к работе данные по белок белковым взаимодействиям у мышей из баз данных STRING, IntAct, mentha, APID, MINT, IID.
- Скачена и подготовлена к работе база данных CORUM по белковым комплексам.
- Скачана и подготовлены к работе информация о белковых комплексах в базе данных IntAct база.
- Установлена биоинформатическая платформа Cytoscape, получены навыки работы с нею. Изучены методы создания Cytoscape плагинов.
- Установлены плагины ClusterONE, MCODE, CytoClaster, SCODE для прогнозирования белковых комплексов, получены навыки работы с ними.