Практическое занятие 3

Ознакомление с методами автоматического анализа текстов (text-mining). Приоритезация генов-кандидатов ресурсом ToppGene. Анализ сверхпредставленных Gene Ontology биологических процессов ресурсом PANTHER.

**Цели занятия:** получение представления (1) о методе автоматического анализа текстов, (2) инструментах для приоритезации генов-кандидатов, (3) ресурсах для поиска сверхпредставленных Gene Ontology биологических процессов.

В ходе практического занятия нужно выполнить описанные ниже упражнения (этапы 1, 2 и 3) и приготовить отчет, включающий:

(а) фамилию,

(б) результаты выполнения упражнений в виде xlsx файла.

Отчет надо отправить на адрес **tiys@bionet.nsc.ru** в приложении к письму в виде xlsx файла.

# *Этап 1. Разработка прототипа системы для извлечения информации о взаимодействиях биологических объектов методом автоматического анализа научных текстов.*

**Цель этапа:** Ознакомление с источниками информации для формирования словарей названий и синонимов биологических объектов. Формирование представления о лингвистических шаблонах, используемых для автоматического анализа текстов. Извлечение информации о взаимодействиях биологических объектов методом автоматического анализа научных текстов. Верификация взаимодействий, извлеченных автоматическими методами.

**Шаг 1. Формирование корпуса статей по тромбозу, которые будут использованы в работе.**

1.1.1. Зайдите на сайт PubMed (<https://www.ncbi.nlm.nih.gov/pubmed>)

1.1.2. Введите в поисковую строку запрос «thrombosis» и нажмите кнопку «Search», в левой части страницы выберите фильтрацию статей за последний год: фильтр «Publication dates», «Custom range...».

1.1.3. В верхнем правом углу страницы выберите «Send to» → «File» → «Format MEDLINE» → «Create File».

**Шаг 2. Формирование словаря названий и синонимов генов человека.**

1.2.1. Зайдите на сайт NCBI Gene (<https://www.ncbi.nlm.nih.gov/gene>)

1.2.2. Пройдите по ссылке и скачайте файл с генами человека: «[Download/FTP](ftp://ftp.ncbi.nih.gov/gene/)» → DATA → GENE\_INFO → Mammalia → Homo\_sapiens.gene\_info.gz

1.2.3. Убедитесь, что в файле «Homo\_sapiens.gene\_info.gz» представлена информация по нескольким подгруппам вида Homo sapiens. Для этого обратите внимание на идентификаторы организма в первом столбце (tax\_id) файла. Введите в поисковую строку идентификаторы этих организмов на сайте базы данных NCBI Taxonomy (<https://www.ncbi.nlm.nih.gov/taxonomy/>) и узнайте, что это за организмы. Удалите из файла «Homo\_sapiens.gene\_info» все строки, не относящиеся к Homo sapiens tax\_id=9606.

1.2.4. Обратите внимание в каких столбцах файла указаны идентификаторы «GeneID», названия и синонимы генов.

**Шаг 3. Формирование списка названий заболевания тромбоз.**

1.3.1. Зайдите на сайт Disease Ontology (<http://disease-ontology.org/>)

1.3.2. Введите в поисковую строку запрос «thrombosis» и нажмите кнопку «Search».

1.3.3. Скопируйте результат поиска в текстовый файл.

**Шаг 4. Создание лингвистических шаблонов, используемых для выявления взаимодействий между биологическими объектами.**

1.4.1. Запишите в текстовый файл не менее 10 примеров ключевых слов, описывающих взаимосвязь между генами и заболеванием тромбоз. Для этого просмотрите статьи из корпуса статей, построенного на **шаге 1**.

Примеры:

is a risk factor

association

polymorphism

allele frequency

mutation

expression

activity

associated

### Шаг 5. Автоматический анализ научных текстов.

1.5.1. Воспользуйтесь скриптом «Text\_mining\_prototype.py» (скопируйте код из [Приложения 1](file:///C:\Users\IgnatevaEV\Downloads\Text_mining_prototype.py#Приложение 1. Код скрипта ) в пустой текстовый файл). Для этого создайте файл start.bat в директории с файлами и напишите туда строку:

python Text\_mining\_prototype.py corpus genes disease templates out\_file, где вместо следующих слов подставьте реальные названия файлов:

corpus — файл с корпусом статей, построенный на **шаге 1**;

genes — файл с названиями и синонимами генов, полученный на **шаге 2**;

disease – файл с названиями заболевания глаукомы, полученный на **шаге 3**;

templates — файл с ключевыми словами, описывающими взаимодействия между биологическими объектами, полученный на **шаге 4**;

out\_file — имя файла, в котором будут результаты.

1.5.2. Запустите файл start.bat двойным кликом (в линукс надо сделать файл start исполняемым [команда в консоли chmod u+x start] и запустить из консоли ./start)

**Шаг 6. Верификация результатов автоматического анализа текстов.**

1.6.1. Импортируйте содержимое выходного файла «out\_file» в новый файл **practice3\_ВашаФамилия.xls**.

1.6.2. Просмотрите 10 выявленных взаимодействий и отметьте в новой колонке таблицы, действительно ли выявленный ген ассоциирован с тромбозом (1-есть ассоциация, 0-нет ассоциации).

1.6.3. Посчитайте точность распознавания взаимодействий в отдельной ячейке:

**Precision = N correct/N total**

**N correct** – число верно распознанных взаимодействий.

**N total** – общее число взаимодействий в тестовом наборе.

**Шаг 7. Формирование отчета по этапу 1**

1.7.1. Переименуйте лист **practice3\_ВашаФамилия.xls** с взаимодействиями в «interactions».

1.7.2. Создайте лист «templates» и скопируйте в него использованные Вами примеры ключевых слов, описывающих взаимосвязь между генами и заболеванием тромбоз, из **шага 4**.

# Этап 2. *Приоритезация ресурсом ToppGene генов-кандидатов**, вовлеченных в ответ клетки на механический стресс (Gene Ontology категория cellular response to mechanical stimulus GO:0071260), к генам, ассоциированных с тромбозом, выявленным на этапе 1.*

**Цель этапа:** Ознакомление с ресурсом ToppGene, позволяющим проводить приоритезацию генов-кандидатов к заданному набору генов.

**Шаг 1. Формирование списка генов-кандидатов, вовлеченных в ответ клетки на механический стресс (Gene Ontology категория cellular response to mechanical stimulus GO:0071260), для приоритезации.**

2.1.1. Откройте в интернет-браузере домашнюю страницу сервиса **Gene Ontology Consortium** (<http://geneontology.org/>[)](http://www.geneontology.org/)/).

2.1.2. На главной странице в поле «Search GO data» введите идентификатор **GO:0071260** соответствующий категории **cellular response to mechanical stimulus**, нажмите «Search».

2.1.3. Нажмите кнопку «Ontology». Кликните на названии анализируемой категории: **сellular response to mechanical stimulus**.

2.1.4. Во вкладке «**Annotations**» нажмите на кнопку «**Organism**» и выберите Homo sapiens, нажав на знак +.

2.1.5. В центре странице нажмите на кнопку «**Download**». Перетащите поле «**Gene/product (bioentity\_label)**» в правую часть окна после поля «**Gene/product (bioentity)**» и нажмите «**Download**». Сохраните результат в текстовый файл и добавьте этот результат к Вашему xls файлу **practice3\_ВашаФамилия.xls** на новый лист под названием «**CRMS**».

**Шаг 2. Приоритезация генов-кандидатов****, вовлеченных в ответ клетки на механический стресс (Gene Ontology категория cellular response to mechanical stimulus GO:0071260), к генам, ассоциированным с тромбозом, выявленным на этапе 1, с помощью ресурса ToppGene.**

2.2.1. Откройте в интернет-браузере домашнюю страницу сервиса ToppGene (<https://toppgene.cchmc.org/>). Выберите ссылку «**ToppGene: Candidate gene prioritization**».

2.2.2. На открывшейся странице выберите «**Entrez ID**» в поле «**Symbol Types**» для окна «**Training Gene Set:**». В качестве тренировочного набора будем использовать гены, ассоциированные с тромбозом, выявленные на этапе 1. Из файла **practice3\_ВашаФамилия.xls** лист **«interactions**» скопируйте идентификаторы GeneID из первой колонки таблицы в окно «**Training Gene Set:**».

2.2.3. В качестве тестового набора будем использовать гены, вовлеченные в ответ клетки на механический стресс GO:0071260. Выберите «**HGNC Symbol**» в поле «**Symbol Types**» для окна «**Test gene set:**». Из файла **practice3\_ВашаФамилия.xls** лист **«CRMS**» скопируйте названия генов из второй колонки таблицы в окно «**Test Gene Set:**».

2.2.4. Нажмите на кнопку «**Отправить**» («**Submit**»). Используйте все настройки по умолчанию. Нажмите кнопку «**Start**».

2.2.5. Когда приоритезация будет закончена (примерно через 0,5-5 минут), нажмите рядом с заголовком «Test Results» кнопку «Download» и скопируйте результат в файл **practice3\_ВашаФамилия.xls** на новый лист «ToppGene».

**Шаг 3. Формирование отчета по этапу 2**

2.3.1. Убедитесь, что в файле **practice3\_ВашаФамилия.xls** представлены следующие листы, полученные в ходе этапа 2:

* «CRMS»
* «ToppGene»

# Этап 3. *Определение сверхпредставленных Gene Ontology биологических процессов для набора генов, ассоциированных с тромбозом, выявленных на этапе 1*.

**Цель этапа:** Ознакомление с ресурсом PANTHER, позволяющим проводить поиск сверхпредставленных Gene Ontology биологических процессов для набора генов.

**Шаг 1. Определение сервисом PANTHER сверхпредставленных Gene Ontology биологических процессов для набора генов, ассоциированных с тромбозом, выявленных на этапе 1.**

3.1.1. Откройте UniProt (<https://www.uniprot.org/>) и перейдите на вкладку Retrieve/ID mapping.

3.1.2. Скопируйте из таблицы ассоциаций генов с тромбозом (из **шага 6 этапа 1**) идентификаторы **GeneID** (первый столбец), вставьте список в поле

**«Provide your identifiers»**. В меню **«Select options»** в выпадающем спике **«From to»** выберите **«Entrez Gene (GeneID)»**. Нажмите **«Submit»**. В появившемся меню **«Filter by»** выберите **«Reviewed»**.Нажмите **«Download»,** в выпадающем списке **«Target list»,** кнопку радио переведите в положение **«uncompressed»** и нажмите **«Go».**

3.1.3. Содержимое загруженного файла скопируйте в свободный столбец на лист **«interactions»** в файл **«practice3\_ВашаФамилия.xls».**

3.1.4. Откройте в интернет-браузере домашнюю страницу Gene Ontology (<http://geneontology.org/>), вставьте список идентификаторов из буфера обмена в текстовое поле под заголовком **«GO Enrichment Analysis»** и нажмите **«Launch»**.

3.1.5. Посмотрите, какие Gene Ontology биологические процессы оказались сверхпредставлены для анализируемого набора генов. Нажмите на кнопку **«Export results»**.

3.1.6. Скопируйте результат на новый лист **«GOEA»** в файл **«practice3\_ВашаФамилия.xls»**.

**Шаг 2. Поиск публикаций, описывающих биологические процессы, вовлеченные в патогенез тромбоза.**

3.2.1. Выберите любой значимо (p-value<0.05 c FDR коррекцией) сверхпредставленный Gene Ontology биологический процесс из **шага 1**.

3.2.2. Найдите в PubMed статью, где описана связь выбранного процесса с тромбозом. Если для выбранного процесса не удалось найти таких публикаций – выберите другой процесс.

3.2.3. Скопируйте название выбранного процесса, PubMed ID и абстракт найденной статьи в в xls-файл **«practice3\_ВашаФамилия.xls»** на новый лист **«BP\_Publ»**.

**Этап 4*\* Статистический анализ экспрессии генов у пациентов с болезнью Альцгеймера с помощью программы MeV.***

**Цель этапа:** Знакомство с программой статистического анализа MeV, использование сверхпредставленности для интерпретации результатов по дифф. экспрессии и подготовка данных для реконструкции ассоциативной генной сети.

**Шаг 1. Загрузка данных в MeV**

4.1.1. Откройте базу данных **«GEO»** и найдите набор **«GDS2601»**.

4.1.2. Скачайте **«DataSet SOFT file»** на странице набора.

4.1.3. Удалите из файла первые строки до строки, начинающейся с **«ID\_REF»**

4.1.4. Откройте программу MeV и в меню file→load data выберете этот файл.

4.1.5. Щелчком левой кнопкой мыши укажите в окне с просмотром первую ячейку со значение экспрессии (всё остальное будет воспринято как заголовки).

**Шаг2. Стат. анализ**

4.2.1. Нажмите на кнопку **«Statistics»** и в выплывающем списке выберите **«Nonparametric tests»**.

4.2.2. Выберите критерий для двухфакторного дизайна эксперимента.

4.2.3. Укажите факторы **«Пол»** и **«Болезнь»** и их градацию. Нажмите **«Next****»**.

4.2.4. В открывшемся окне перетащите названия экспериментов в соответствующие ячейки таблицы (посмотреть соответствие можно на страничке набора **«GDS2601»** вGEO).

4.2.5. После расчёта статистики, найдите в иерархии слева результат по дифференциально экспрессирующимся генам болезни Альцгеймера. Выберите этот пункт.

4.2.6. Правой кнопкой мыши щёлкните по графическому изображению экспрессии и в выпашем списке выберите **«Save cluster...»**

4.2.7. Сохраните результат в **«practice3\_ВашаФамилия.xls»** на лист **«Alz\_diff»**

**Шаг 3. Поиск сверхпредставленных процессов**

4.3.1. Откройте в интернет-браузере домашнюю страницу Gene Ontology (http://geneontology.org/), вставьте список подходящих идентификаторов из сохранённого кластера в текстовое поле под заголовком «**GO Enrichment Analysis**» и нажмите «**Launch**».

4.3.2. Посмотрите, какие Gene Ontology биологические процессы оказались сверхпредставлены для анализируемого набора генов. Нажмите на кнопку «**Export results**».

4.3.3. Скопируйте результат на новый лист «**GOEA\_Alz\_diff**» в файл **«practice3\_ВашаФамилия.xls»**.

4.3.4. Предположите с чем может быть связана Болезнь Альцгеймера и найдите в PubMed есть ли статьи на эту тему, когда наблюдался пик числа статей по этой гипотезе. Напишите какой запрос вы сделали в PubMed и укажите год максимального интереса по этой теме на новом листе **«Conclusion»** файла **«practice3\_ВашаФамилия.xls»**.

### Приложение 1. Код скрипта «Text\_mining\_prototype.py», язык программирования Python [(вернуться к шагу 5)](#_toc55)

import sys

import re

corpus=sys.argv[1]

genes=sys.argv[2]

disease=sys.argv[3]

templates=sys.argv[4]

out\_file=sys.argv[5]

def foo(corpus, genes, disease, templates, out\_file):

f=open(corpus)

A=[]

while 1:

line=f.readline()

if line=="":break

else:

A.append(line)

A="".join(A)

A=A.split("\n\n")

AD={}

for a in A:

if "AB - " in a:

k=a.split("PMID- ")[1].split("\n")[0]

s1=a.split("\n")

s2=" ".join(s1)

s3=s2.split("AB - ")[1]

s4=s3.split(" - ")[0]

AD[k]=s4

f=open(genes)

f.readline()

exceptions=["abeta", "best", "beta2", "cell", "cord", "crest", "damage", "delta", "face", "fact", "find", "flash", "great", "held", "homeobox", "impact", "iris", "large", "light", "mask", "mass", "mice", "minor", "pick", "pilot", "pole", "post", "pars", "past", "race", "scar", "simple", "spatial", "step", "stop", "trade", "traits", "tube"]

G={}

while 1:

line=f.readline().lower()

if line=="":break

else:

gene\_id=line.split("\t")[1]

n1=line.split("\t")[2]

if len(n1)>3:

if n1 not in exceptions:

G[n1]=gene\_id

n2=line.split("\t")[4]

if "|" in n2:

n3=n2.split("|")

for n4 in n3:

if len(n4)>3:

if n4 not in exceptions:

G[n4]=gene\_id

else:

if n2!="":

if len(n2)>3:

if n2 not in exceptions:

G[n2]=gene\_id

f=open(disease)

DN=[]

while 1:

line=f.readline()

if line=="":break

else:

DN.append(line[:-1].split("\t")[1])

f=open(templates)

T=[]

while 1:

line=f.readline()

if line=="":break

else:

T.append(line[:-1])

out=open(out\_file,"w")

out.write("GeneID\tGene name or synonym\trelation\tdisease name\tPubMed ID\tsentence\n")

for k in AD.keys():

sents0=re.sub( '\s+', ' ', AD[k]).strip()

sents=sents0.split(". ")

for s in sents:

for t in T:

if t in s.lower():

for d in DN:

if d in s.lower():

words=s.split(" ")

for w in words:

try:

gene=G[w.lower()]

out.write(gene+"\t"+w+"\t"+t+"\t"+d+"\t"+k+"\t"+s+"\n")

except:pass

out.close()

if \_\_name\_\_ =='\_\_main\_\_':

args = sys.argv[1:]

if len(args)!=5:

print "usage:python Text\_mining\_prototype.py corpus genes disease templates out\_file"

sys.exit(0)

else:

foo(args[0], args[1], args[2], args[3], args[4])