# Отчёт по Лабораторной № 2

Выполнил студент 2 курса СПбАУ Есиков Сергей

## 1. PubMed

**Задача:** Составить список университетов, в которых работают соавторы Pavel Pevzner.

### Метод решения

1. Поиск всех научных работ

```
    handle = Entrez.esearch(
        db="pubmed",
        term=f"{author_name}[Author]",
        retmax=500,
        usehistory="y"
    )
    record = Entrez.read(handle)
    count = int(record["Count"])
    webenv = record["WebEnv"]
    query_key = record["QueryKey"]
    handle.close()
```

```
handle = Entrez.efetch(
        db="pubmed",
        rettype="xml",
        retmode="xml",
        webenv=webenv,
        query_key=query_key,
        retstart=start,
        retmax=batch_size
    )
    data = handle.read().decode('utf-8')
    handle.close()
```

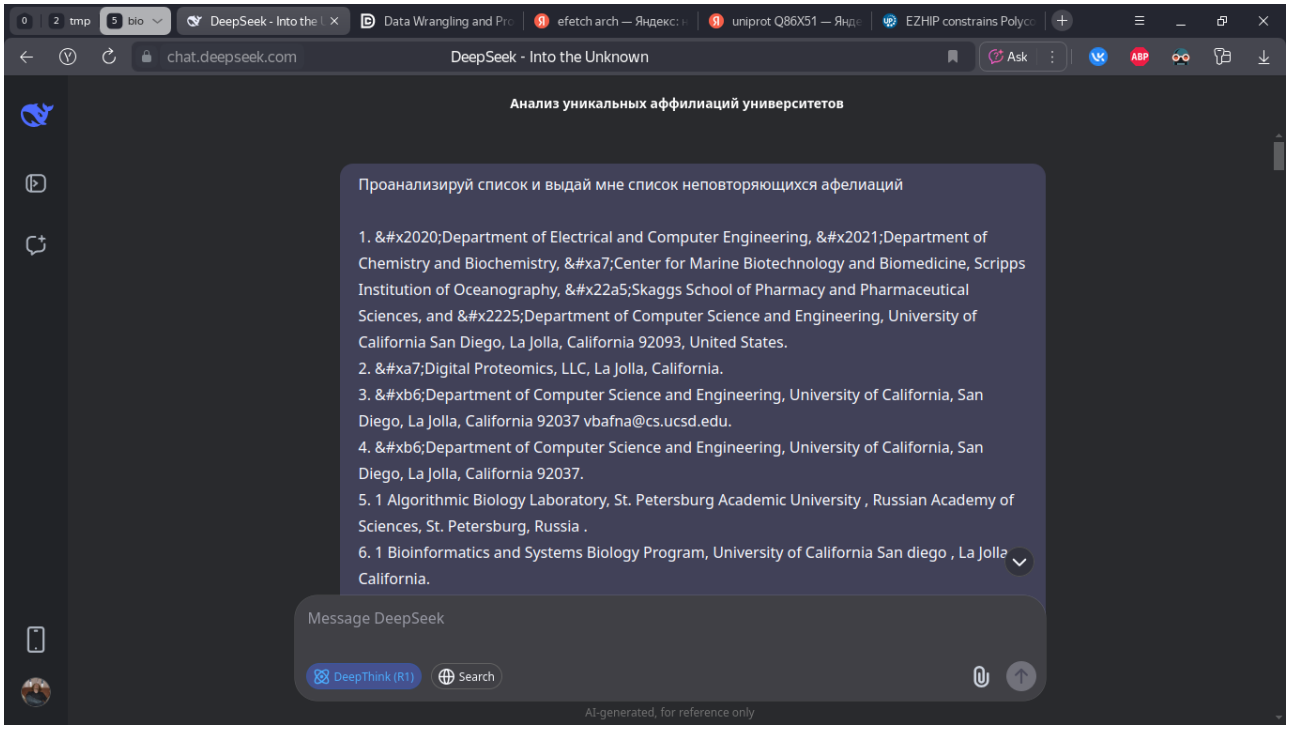
2. Извлечение аффилиаций по тегам

```
# Метод 1: Поиск стандартных тегов Affiliation
aff_pattern = r'<Affiliation[^>]*>([^<]+)</Affiliation>'
matches = re.findall(aff_pattern, xml_data)

if matches:
    return {clean_affiliation(m) for m in matches}
```

```
    # Метод 2: Резервный поиск в формате Medline
    ad_pattern = r'AD\s*-\s*([^\n]+)'
    matches = re.findall(ad_pattern, xml_data)
    return {clean_affiliation(m) for m in matches}
```

3. Фильтрация потенциальных потворений




## Результаты

- Количество уникальных аффилиаций после парсинга: 686
- Количество университетов после доп обработки: 81
- Список университетов в Алфавитном порядке:
    1. Academia Sinica
    2. Amazon
    3. Arima Genomics
    4. Ascus Biosciences
    5. Australian Centre for Ecogenomics
    6. Baylor College of Medicine
    7. California Institute of Technology
    8. Carnegie Mellon University
    9. Cedars-Sinai Medical Center
    10. Clemson University
    11. Cornell University
    12. DNAnexus
    13. Digital BioLogic
    14. Duke University
    15. ETH Zurich
    16. Ecole Polytechnique Fédérale de Lausanne
    17. Emory University

18. Francis Crick Institute
19. Harvard University
20. Helmholtz Institute
21. Howard Hughes Medical Institute
22. Illumina
23. Jackson Laboratory
24. Johns Hopkins University
25. Massachusetts Institute of Technology
26. Max Planck Institute
27. Misvik Biology Ltd
28. National Geographic Society
29. National Institutes of Health
30. National Sun Yat-sen University
31. Pacific Biosciences
32. Pacific Northwest National Laboratory
33. Phase Genomics
34. Rockefeller University
35. San Diego Zoo Wildlife Alliance
36. Scripps Research Institute
37. Senckenberg Research Institute
38. Sirenas Marine Discovery
39. Smithsonian Tropical Research Institute
40. Stanford University
41. St. Petersburg Academic University
42. St. Petersburg State University
43. Thermo Fisher Scientific
44. University of Alabama
45. University of Bari
46. University of Birmingham
47. University of British Columbia
48. University of California, Berkeley
49. University of California, Santa Cruz
50. University of California San Diego
51. University of Chicago
52. University of Chicago Marine Biological Laboratory
53. University of Colorado Denver
54. University of Cambridge
55. University of Florida
56. University of Geneva
57. University of Hamburg
58. University of Hawaii at Manoa
59. University of Illinois Chicago
60. University of Lorraine
61. University of Michigan
62. University of Minnesota
63. University of North Carolina Chapel Hill

64. University of Notre Dame
65. University of Oklahoma
66. University of Oregon
67. University of Pennsylvania
68. University of Pittsburgh
69. University of Porto
70. University of Queensland
71. University of Regensburg
72. University of Rennes 1
73. University of São Paulo
74. University of Southern California
75. University of Tennessee Health Science Center
76. University of Texas A&M
77. University of Virginia
78. University of Washington
79. University of Western Australia
80. University of Wisconsin-Madison
81. University of Tartu

---

## 2. DB-NCBI-PubMed

**Задача:** Посчитать публикации в PubMed за период [Дата начала] - [Дата конца] и получить их PMC ID.

### Метод решения

Поиск осуществляется через прямой запрос к `eutils` по отдельности на каждый день из запрошенных

Далее после получения ID всех статей идёт поиск тех из них что обладают PMC ID

Извлечение PM ID и PCM ID происходит с помощью xml парсинга по тегам

- сам скрипт
- PMC ID

### Результаты

```
Всего найдено PM ID: 26286
Всего найдено PMC ID: 15118
Результаты сохранены в pmc_list.txt
```

---

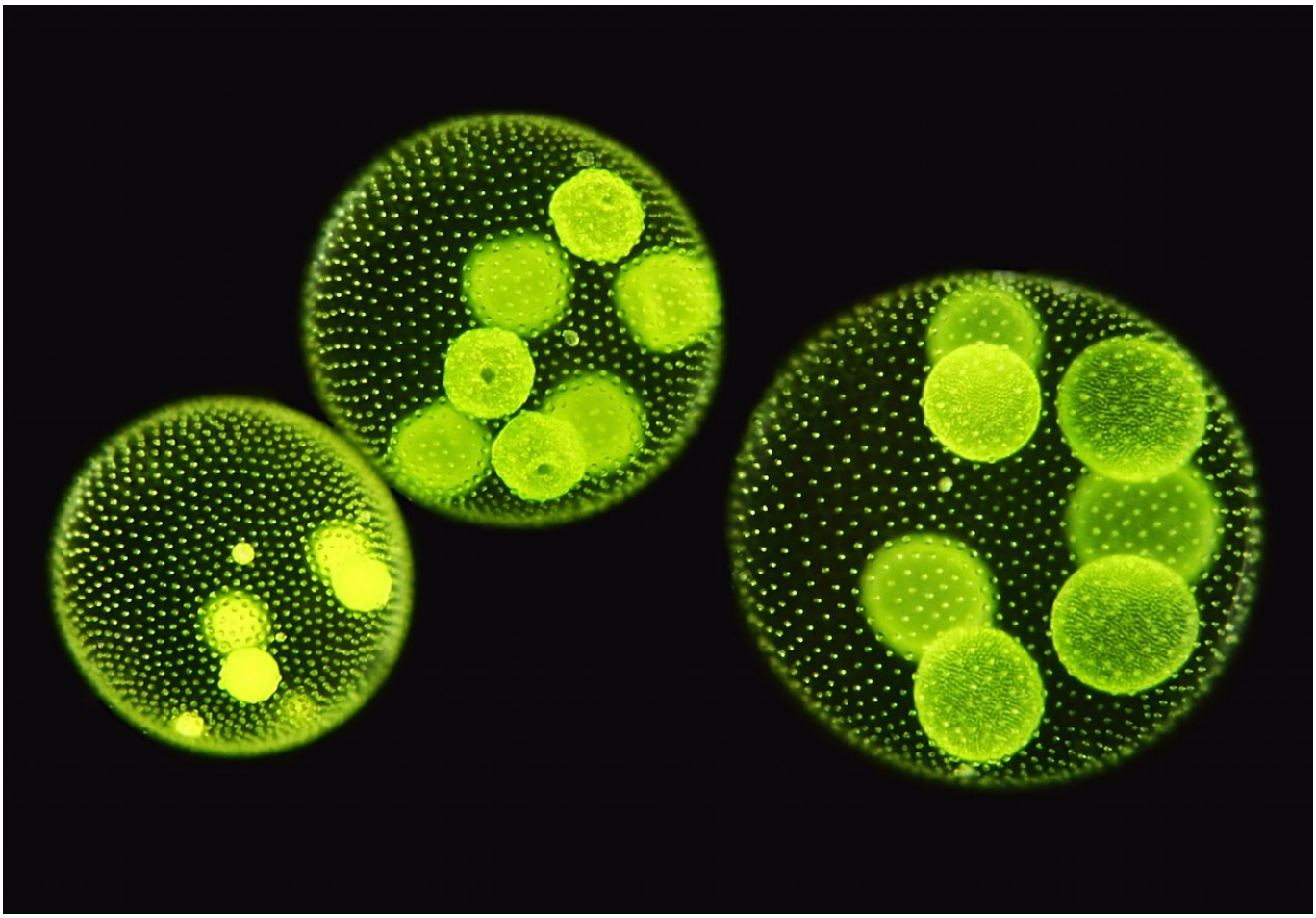
## 3. DB-NCBI-volvox

**Задача:** Скачать кодирующие последовательности (CDS) для организма Volvox в формате FASTA.

### Описание организма

**Volvox** — род пресноводных колониальных зелёных водорослей (*Chlorophyta*), представляющий важную модель в исследованиях эволюции многоклеточности.

Колонии Volvox демонстрируют примитивную форму "разделения труда" между клетками — ключевой этап эволюции сложных организмов.



## Метод решения

Обращаемся к `DB_NCBI` с помощью NCBI E-Utilities

- Установка:

```
& sh -c "$(curl -fsSL
https://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/install-edirect.sh)"
```

- Сам запрос:

```
& esearch -db nucleotide -query "Volvox[Organism] AND (cds[Feature] OR
coding[Title])" | \
efetch -format fasta > volvox_cds.fasta
```

- Пояснения:

1. Volvox[Organism] записи для организма Volvox.
2. cds[Feature] последовательности с аннотированными кодирующими регионами (CDS).
3. coding[Title] записи, где в заголовке есть слово "coding"
4. Далее форматирование в `FASTA` и вывод в файл

### Результаты

- Файл с данными: volvox_cds.fasta
- Количество последовательностей:

```
& cat ./volvox_cds.fasta | grep '>' | wc -l
03
```

---

## 4. DB-Ensembl

**Задача:** Найти список хромосом и их длину для последнего референса генома мыши.

### Метод решения

Прямой запрос к `Ensembl REST API` с последующей филтрацией json выдачи

1. команда

```
& curl "https://rest.ensembl.org/info/assembly/mus_musculus?content-
type=application/json" | jq '.top_level_region[] |
select(.coord_system == "chromosome") | {name, length}' >> output
```

   - сервер

     `https://rest.ensembl.org/info/assembly/mus_musculus`

   - вывод данных

     `content-type=application/json`

2. парсинг json через утилиту `jq`

   - Начальный корень

     `.top_level_region[]`

   - Выборка по атрибуру coord_system представляющему хромосому

     `select(.coord_system == "chromosome")`

   - Фильтр результата до имени хромосомы и её размера

```
{name, length}
```

3. Вывод в файл

```
>> output
```

### Результаты

| | | | |
|---|---|---|---|
| 1: 195154279 | 2: 181755017 | 3: 159745316 | 4: 156860686 |
| 5: 151758149 | 6: 149588044 | 7: 144995196 | 8: 130127694 |
| 9: 124359700 | 10: 130530862 | 11: 121973369 | 12: 120092757 |
| 13: 120883175 | 14: 125139656 | 15: 104073951 | 16: 98008968 |
| 17: 95294699 | 18: 90720763 | 19: 61420004 | |

---

## 5. DB-UNIPROT

**Задача:** Найти человеческие белки, содержащие подпоследовательность [Аминокислотная последовательность].

### Метод решения

Парсинг всех пептидов доступных на ресурсе и последующее их выравнивание на искомую подпоследовательность

1. Парсинг пептидов

```
curl 'https://rest.uniprot.org/uniprotkb/stream?
fields=accession,sequence&format=tsv&query=(*)+AND
+(model_organism:9606)' > allProteins.txt
```

   - `https://rest.uniprot.org/uniprotkb/stream`

     Сервер

   - `fields=accession,sequence`

     Поля вывода

   - `format=tsv`

     Формат вывода

   - `query=(*)+AND+(model_organism:9606)`

     Параметры запроса

2. Выравнивание

- Поиск включений подстроки дал нулевой результат, так что было принято решение допустить погрешность в 1 букву

1. Для этого поиск вхождений был заменен на выравнивание подстроки по белку со следующими параметрами

    - Стоимость совпадения: положительное число `n`
    - Стоимость несовмадения: 0
    - Стоимость любых решений в сторону гепа: сильно отрицательное число, сравнимо большее `n`

```
from Bio import Align
from Bio.Seq import Seq

pepride = Seq("AAVGPQKATA")
aligner = Align.PairwiseAligner()
aligner.mode = 'local'
aligner.match_score = 2
aligner.mismatch_score = 0
aligner.open_gap_score = -100
aligner.extend_gap_score = -100
```

2. Далее выравнивание и выбор наилучшего

```
with open("./allProteins.txt") as f:
    i = 1
    for line in f:
        note = line.split()
        alignments = aligner.align(Seq(note[1]), pepride)
        best_alignment = alignments[0] if alignments else None
```

3. Если `score` наилучшего совпадения меньше чем

    - Стоимость совпадения
    - Умноженная на
    - Длину подстроки за вычетом количества допустимых ошибок

    Значит выравнивание либо слишком грязное либо вообще отсутствует

```
if (best_alignment) and (best_alignment.score >= 2 * (len(pepride) -
K)):
    print(best_alignment)
    res.write(f"{note[0]}, {note[1]},
{str(best_alignment.coordinates[0])}\n")
else:
    print(i, best_alignment.score if best_alignment else -1, end='\r')
i += 1
```

## Результаты

На выходе получилоась два результата, но они одинаковые

EZHIP constrains Polycomb Repressive Complex 2 activity in germ cells.

| Имя белка | Длина (a.o.) | UniProt ID |
|---|---|---|
| EZHIP | 503 | Q86X51 |

Если допустить три ошибка, выборка сильно увеличивается (с двумя ошибками резуьтат не отчиается от первого)

| UniProt ID | Длина (a.o.) |
|---|---|
| O60755 | 368 |
| P13598 | 275 |
| P23471 | 2315 |
| P41225 | 446 |
| Q86X51 | 503 |
| Q8NA72 | 575 |
| Q9NPD3 | 245 |
| Q9NYZ4 | 499 |
| Q9Y5L0 | 923 |
| A0A494C036 | 1513 |
| A0A494C055 | 2347 |
| A0A494C087 | 2354 |
| A0A494C0U4 | 2059 |
| A0A494C1B4 | 1908 |
| A0A494C1H9 | 1494 |
| A0A494C1R4 | 1487 |
| A0A494C1J5 | 698 |
| B3KN17 | 530 |
| C9JT30 | 390 |
| D6RDG4 | 547 |
| E9PI41 | 261 |

| UniProt ID | Длина (a.o.) |
|---|---|
| J3QRT5 | 251 |
| Q6FHE2 | 275 |
| A0A0D9SET9 | 83 |
| A0A515VFR0 | 503 |
| A0A994J4D9 | 302 |
| A8KAP5 | 275 |
| B2R6H7 | 923 |
| B3KM69 | 523 |
| B3KMX1 | 480 |
| B4DFE7 | 808 |
| C9J7E5 | 957 |
| E9PFH4 | 857 |
| H0Y880 | 109 |
| J3QKR4 | 174 |
| J3QQX6 | 214 |
| J3QRQ1 | 216 |
| Q59ED3 | 344 |
| Q6L9N7 | 108 |
| Q6NXN2 | 316 |
| Q6WG70 | 152 |

количество: 41

...

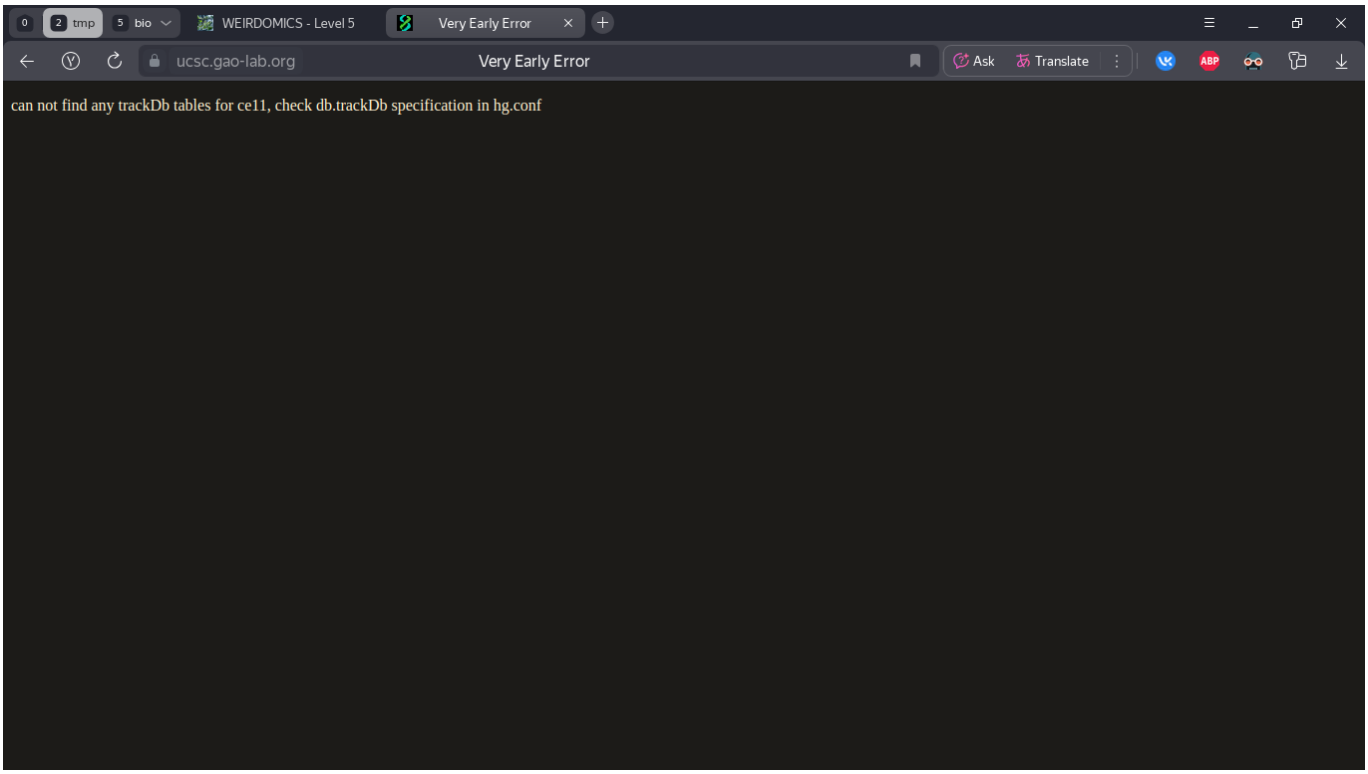
---

# 6. DB-UCSC

**Задача:** Извлечь последовательность генома кошки (хромосома [Имя], позиции [Start]-[End]).

## Метод решения

К сожалению в процессе я столкнулся с проблемой в которой так и не смог разобраться А именно `UCSC Tablet Browser` категорически отказывался позволять мне сменить датасет на млекопитающих


0
2 tmp
5 bio
WEIRDOMICS - Level 5
Very Early Error
ucsc.gao-lab.org
Very Early Error
Ask
Translate
can not find any trackDb tables for ce11, check db.trackDb specification in hg.conf