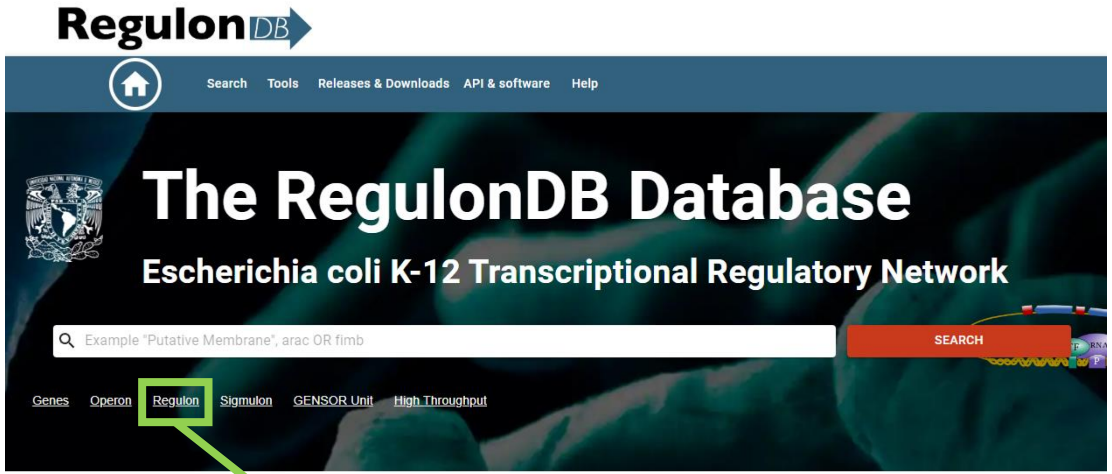
Molekulare Daten von Bakterien in einen Knowledge Graph überführen

DIS18 Projektarbeit I - Gruppe 4

Kerem & Tim

31.07.2024

Ziel



Regularprotein

Regulon

DicF	repressor	ftsZ
DicF	repressor	xylR
DicF	repressor	pykA
DicF	repressor	manX

Gene

WIKIDATA

Schritt 1 – Daten sammeln

1. Regulon Kennzeichnung

RegulonDB

Search Tools Releases & Downloads API & softwa

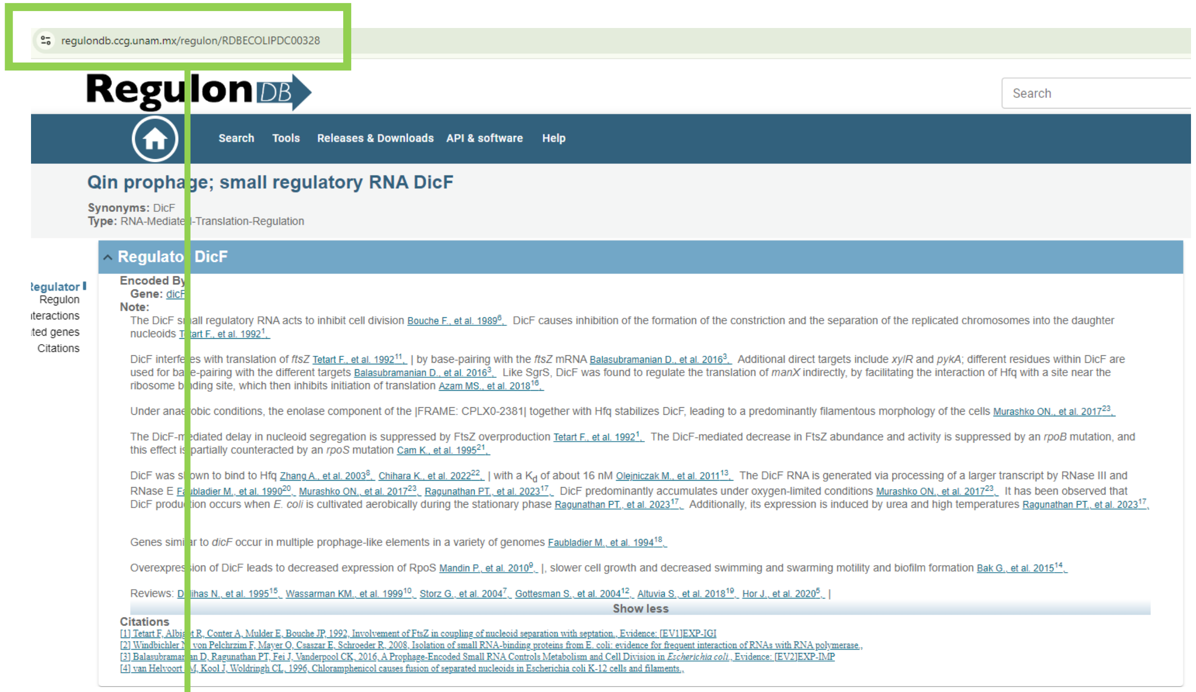
Regulons

Regulon	Regulated		
Name	Genes	Operons	Transcription Unit
Filter... (287)	Min (1) Max (57	Min Max (24	Min Max (31
	308	141	169
DicF	4	0	0
DsrA	10	4	4
GcvB	60	16	17
McaS	6	2	3

DicF, DsrA, GcvB, McaS, …

+

2. Link zum DB Eintrag



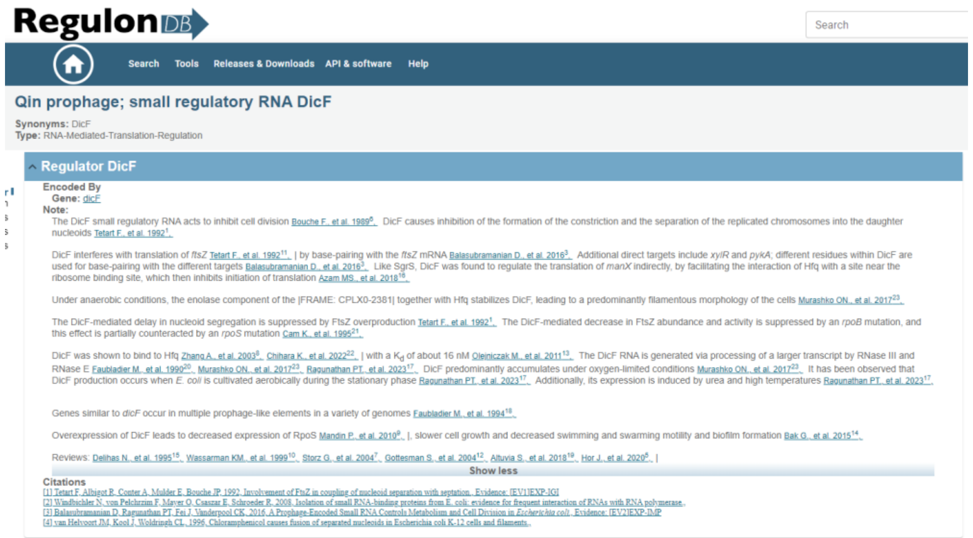
regulondb.ccg.unam.mx/regulon/RDBECOLIPDC00328

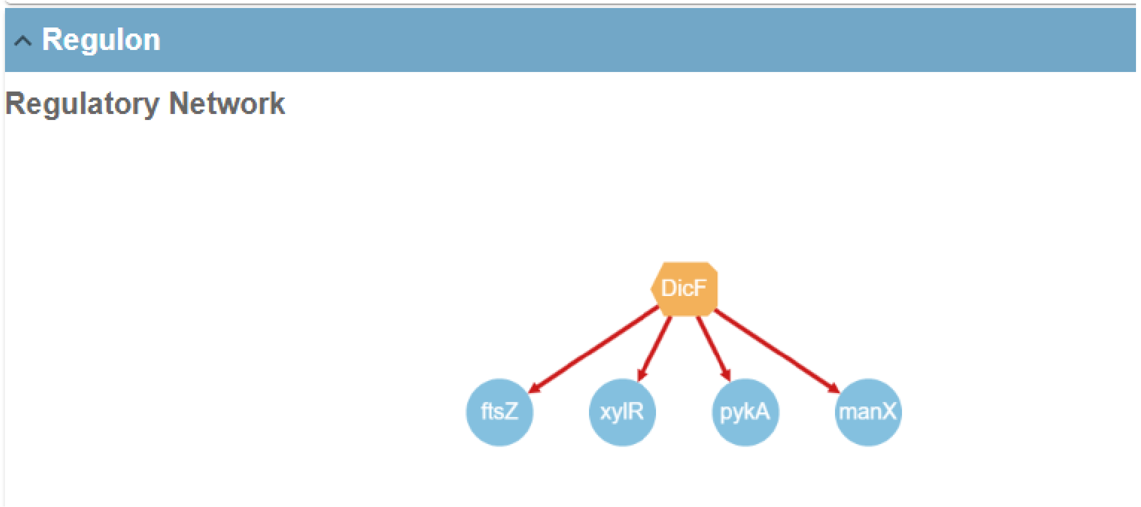
Schritt 1 – Daten sammeln

```python
chrome_options = webdriver.ChromeOptions()
#chrome_options.add_argument("--headless")
chrome_options.add_argument("--no-sandbox")
chrome_options.add_argument('--window-size=1920,1080')
chrome_options.add_experimental_option("prefs", {
    "download.prompt_for_download": False,
    "download.directory_upgrade": True,
    "safebrowsing.enabled": True
})
driver = webdriver.Chrome(options=chrome_options)
actions = ActionChains(driver)

def click_button_wait(XPATH):
    WebDriverWait(driver, 10).until(EC.element_to_be_clickable((By.XPATH, XPATH))).click()

def find_next_page_button():
    driver.execute_script("window.scrollTo(0, document.body.scrollHeight);") # scroll down
    button_xpath = "/html/body/div[1]/div/div[2]/div/div[2]/div/div[3]/div/table/tbody/tr[19]/td/div/div/div/button[3]"
    button = WebDriverWait(driver, 10).until(EC.element_to_be_clickable((By.XPATH, button_xpath)))
    button.click()

# seite aufrufen
driver.get("https://regulondb.ccg.unam.mx/regulon")

# Cookies Rejecten
click_button_wait('/html/body/div[1]/div/div[4]/div/div[3]/button[2]')


is_last_page = False
regulon_links = {}
try:
    while True:
        # Get Table Body
        table_body = WebDriverWait(driver, 10).until(EC.visibility_of_element_located((By.XPATH,'/html/body/div[1]/div/di
        # jedes tr (zeile der tabelle) sammeln
        elements = table_body.find_elements(By.XPATH,".//*[contains(@id, 'tr')]")

        for i in range(0, len(elements)):
            try:
                link = elements[i].find_elements(By.XPATH,".//*")[0].find_element(By.XPATH,"div/div/a").get_attribute("hr
                name = elements[i].find_elements(By.XPATH,".//*")[0].find_element(By.XPATH,"div/div/a").get_attribute("te
                regulon_links[name] = link
            except:
                pass

        find_next_page_button()
except:
    print("finished")
```

```
{'DicF': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC00328',
 'DsrA': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC00358',
 'GcvB': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC03275',
 'McaS': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC03411',
 'GadY': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC03413',
 'MicF': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC03508',
 'OxyS': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC03909',
 'ChiX': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04122',
 'SgrS': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04132',
 'IstR-1': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04134'
 'RseX': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04142',
 'RydC': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04143',
 'OhsC': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04144',
 'SymR': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04148',
 'MgrR': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04149',
 'FnrS': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04150',
 'ArrS': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04151',
 'SdsN': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04154',
 'AgrB': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04160',
```

Schritt 2 – Daten abfragen



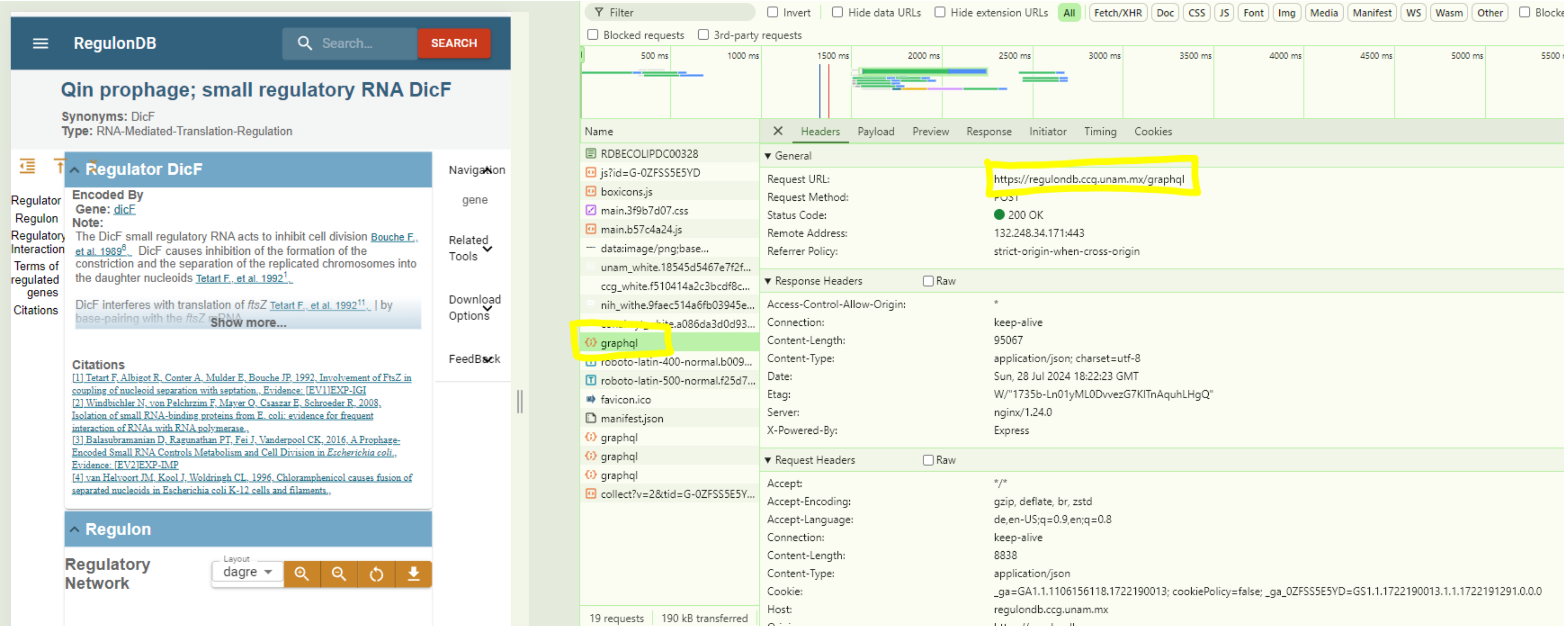


Tripletts

Regulatory Interactions

Active Conformation Name	Function	Regulated Entity Name	Regulated Entity Type	Distance to First Gene	Distance to Promoter	Regulated Genes	RBS LeftEndPosition
Filter... (1)	Filter... (1)	Filter... (4)	Filter... (1)	Min (-15 Max (42	Filter... (1)	Filter... (4)	Min (105277) Max (373498
DicF	repressor	ftsZ	gene	-15		ftsZ	105277
DicF	repressor	xylR	gene	13		xylR	3734988
DicF	repressor	pykA	gene	-10		pykA	1937631
DicF	repressor	manX	gene	42		manX	1902082

Schritt 2 – Daten abfragen



Schritt 2 – Daten abfragen

```
url = "https://regulondb.ccg.unam.mx/graphql"
```

Graphql Endpoint

```
request1["variables"]["advancedSearch"] = "RDBECOLIPDC00328[_id]"
response = json.loads(requests.post(url = url, json = request1, verify = False).content)
```

```
C:\Anaconda\lib\site-packages\urllib3\connectionpool.py:1061: InsecureRequestWarning: Unverified HTTPS request is being made to host 'regulondb.ccg.unam.mx'. Adding certificate verification is strongly advised. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-warnings
  warnings.warn(
```

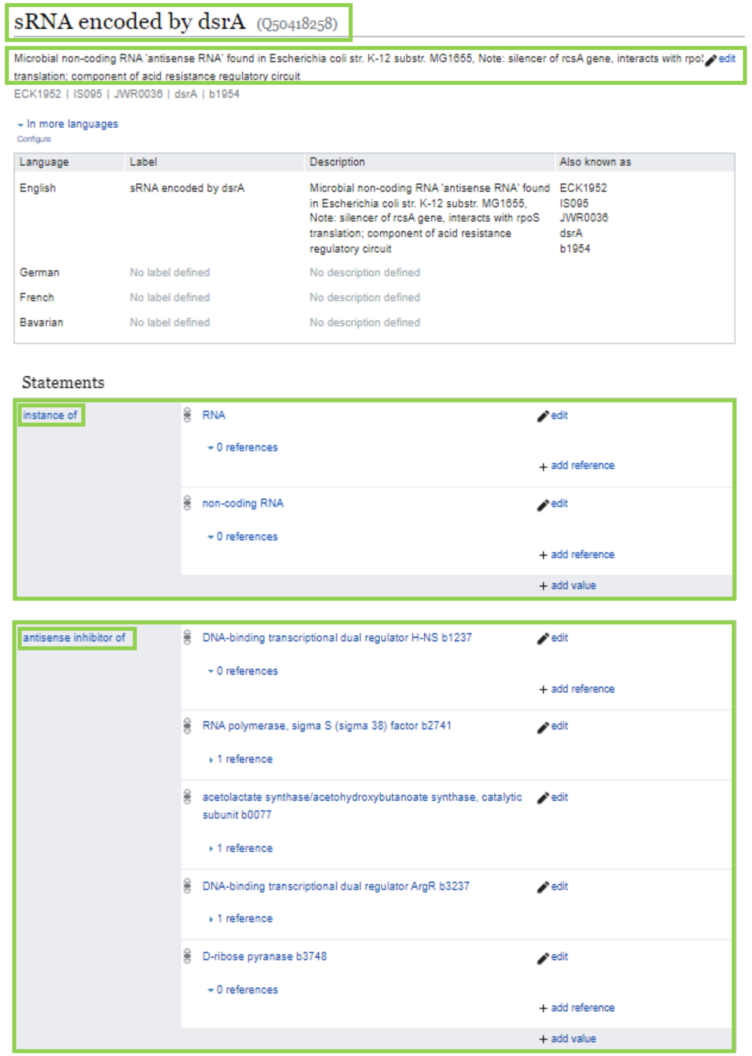
```
response
```

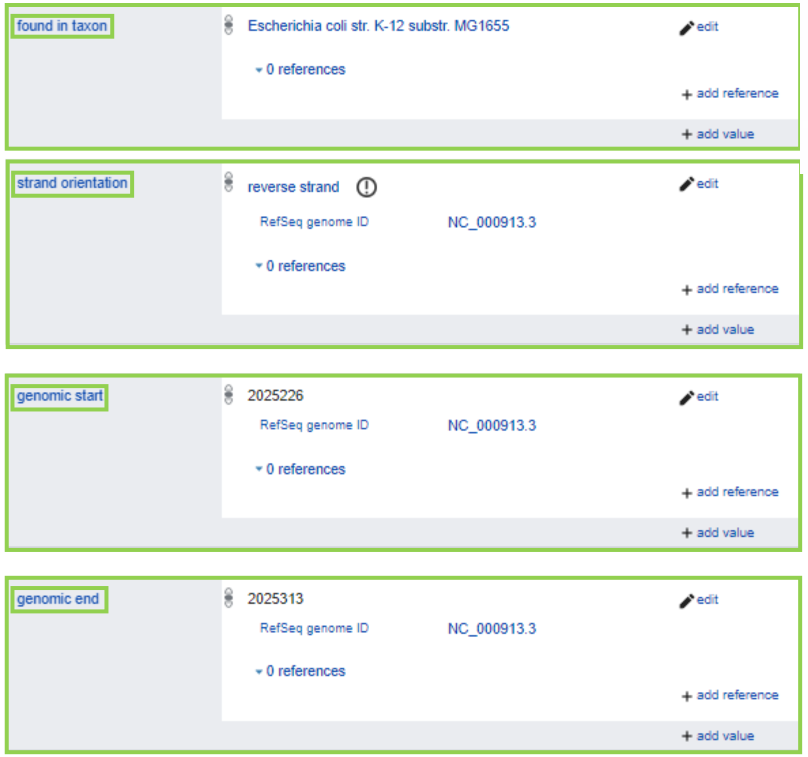

```
{'data': {'getRegulonBy': {'data': [{'_id': 'RDBECOLIPDC00328',
     'aligmentMatrix': {'aligment': None,
      'consensus': None,
      'matrix': None,
      'urlMatrixQualityResult': None,
      'urlPWMLogo': None,
      '__typename': 'AligmentMatrix'},
     'allCitations': [{'evidence': {'_id': 'RDBECOLIEVC00062',
        'additiveEvidenceCodeRule': None,
        'code': 'EXP-IGI',
        'name': 'Inferred from genetic interaction',
        'type': 'W',
        '__typename': 'Evidence'},
       'publication': {'_id': 'RDBECOLIPRC06277',
        'authors': ['Tetart F',
         'Albigot R',
         'Conter A',
         'Mulder E',
         'Bouche JP'],
        'citation': 'Tetart F, Albigot R, Conter A, Mulder E, Bouche JP, 1992, Involvement of FtsZ in coupling of nucleoid separation with septation.',
        'pmid': '1552861',
        'title': 'Involvement of FtsZ in coupling of nucleoid separation with septation.',
```



```
{'DicF': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC00328',
 'DsrA': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC00358',
 'GcvB': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC03275',
 'McaS': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC03411',
 'GadY': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC03413',
 'MicF': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC03508',
 'OxyS': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC03909',
 'ChiX': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04122',
 'SgrS': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04132',
 'IstR-1': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04134'
 'RseX': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04142',
 'RydC': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04143',
 'OhsC': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04144',
 'SymR': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04148',
 'MgrR': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04149',
 'FnrS': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04150',
 'ArrS': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04151',
 'SdsN': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04154',
 'AgrB': 'https://regulondb.ccg.unam.mx/regulon/RDBECOLIPDC04160',
```

Schritt 3 – Daten Preparen





UPLOAD REGULONS

1. kurze beschreibung des items
2. konstant: instance of [P31] (RNA [Q11053] / non-coding RNA [Q427087])
3. konstant: found in taxon [P703] (Escherichia coli str. K-12 substr. MG1655 [Q21102938])
4. statements für gene: inhibitor / repressor / dual
5. strand orientation [P2548] (reverse [Q22809711] / forward [Q22809680])
6. genomic start [P644] / genomic end [P645]

```
list(regulon_data.values())[0]
```

```
{'beschreibung': 'The DicF small regulatory RNA acts to inhibit cell division.',
 'left_position': 1649382,
 'right_position': 1649434,
 'strand_orientation': 'Q22809711'}
```

```
list(triplets.values())[0]
```

```
[{'subjekt': 'DicF', 'predikat': 'repressor', 'objekt': 'ftsZ'},
 {'subjekt': 'DicF', 'predikat': 'repressor', 'objekt': 'xylR'},
 {'subjekt': 'DicF', 'predikat': 'repressor', 'objekt': 'pykA'},
 {'subjekt': 'DicF', 'predikat': 'repressor', 'objekt': 'manX'}]
```

```
print(gene_ids["ftsZ"])
print(gene_ids["xylR"])
print(gene_ids["pykA"])
print(gene_ids["manX"])
```

```
Q127734188
Q127735541
Q127735563
Q127735585
```

Schritt 4 – Daten hochladen (PyWikiBot)

<u>Genes erzeugen</u>

```python
i = 0
for key, value in list(gene_descriptions.items())[6:]:
    label = key
    description = value

    # Create Item
    data = {
        'labels': {'en': {'language': 'en', 'value': label}},
        'descriptions': {'en': {'language': 'en', 'value': str(description)}}
    }
    new_item = pywikibot.ItemPage(repo)
    new_item.editEntity(data, summary=f'Creating new item: {label}')

    # Create Statement
    claim = pywikibot.Claim(repo, "P31")  # instance of
    target = pywikibot.ItemPage(repo, "Q7187")  # gene
    claim.setTarget(target)
    new_item.addClaim(claim, summary='Adding education information')

    gene_ids[key] = new_item.getID()
    print(key, i, "/", len(gene_descriptions))
    i += 1
```

```
Sleeping for 9.4 seconds, 2024-07-24 15:43:27
citR 2686 / 2697
Sleeping for 9.3 seconds, 2024-07-24 15:43:38
Sleeping for 9.6 seconds, 2024-07-24 15:43:47
glsB 2687 / 2697
Sleeping for 9.4 seconds, 2024-07-24 15:43:57
Sleeping for 9.3 seconds, 2024-07-24 15:44:08
yneG 2688 / 2697
Sleeping for 9.2 seconds, 2024-07-24 15:44:18
Sleeping for 9.4 seconds, 2024-07-24 15:44:27
sfmC 2689 / 2697
Sleeping for 9.4 seconds, 2024-07-24 15:44:37
Sleeping for 9.5 seconds, 2024-07-24 15:44:47
```

RegulonDB

Search Tools Releases & Downloads API & software Help

Genes

HIDE COLUMNS DOWNLOAD

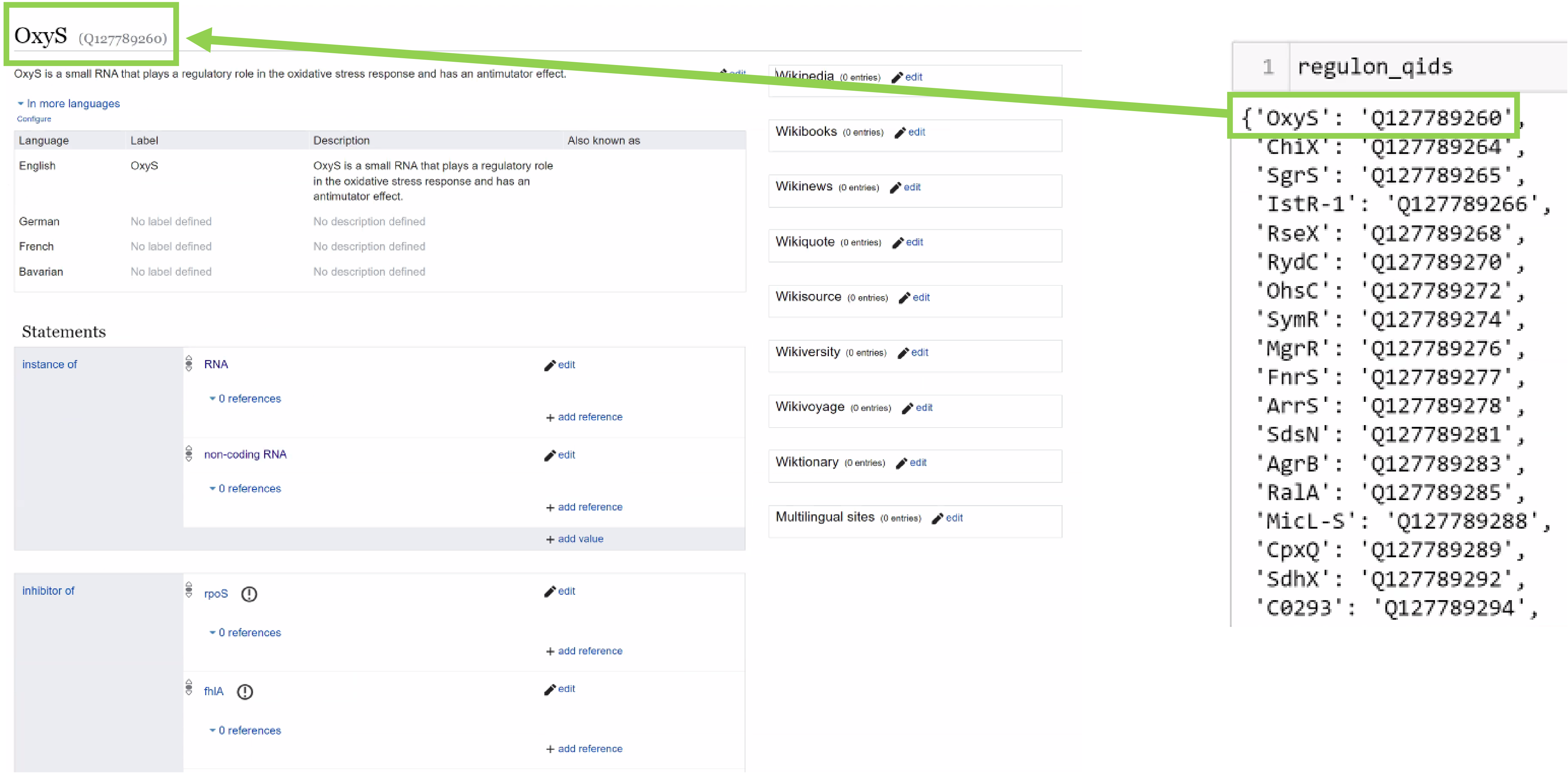
Name	Synonyms	Product(s)
Filter... (4747)	Filter... (4748)	Filter... (4521)
3'ETS-*leuZ*	G0-16636	small regulatory RNA 3'ETS*leuZ*
C0067	G0-8896	
C0293	G0-8897	e14 prophage; small regulatory RNA C0293
C0299	G0-8898	
C0362	G0-8902	
C0465	G0-8903	
C0614	G0-8907	
C0664	G0-8911	
C0719	G0-8912	
G0-10697	G0-10697	
G0-10698	G0-10698	
G0-10700	G0-10700	
G0-10702	G0-10702	
G0-10703	G0-10703	

Schritt 4 – Daten hochladen (PyWikiBot)

```python
def generate_claims(regulon, info):
    label = regulon
    if info["beschreibung"] != None:
        description = info["beschreibung"]
    else:
        description = "description: " + regulon

    if len(str(description)) >= 250:
        description = str(description[:249])
    # Create Item
    data = {
        'labels': {'en': {'language': 'en', 'value': label}},
        'descriptions': {'en': {'language': 'en', 'value': str(description)}}
    }
    new_item = pywikibot.ItemPage(repo)
    new_item.editEntity(data, summary=f'Creating new item: {label}')

    claims = []
    # claims für instance of
    instance_of_rna = {'mainsnak': {'snaktype': 'value', 'property': "P31", 'datavalue': {'value': {'entity-type': 'item', 'numeric-id': "11053"}, 'type': 'wikibase-entityid'}}, 'type': 'statement', 'rank': 'normal'}
    claims.append(instance_of_rna)
    instance_of_non_coding_rna = {'mainsnak': {'snaktype': 'value', 'property': "P31", 'datavalue': {'value': {'entity-type': 'item', 'numeric-id': "427087"}, 'type': 'wikibase-entityid'}}, 'type': 'statement', 'rank': 'normal'}
    claims.append(instance_of_non_coding_rna)

    # claims für funktionen der gene (repressor, activator, dual)
    for i in range(0, len(triplets[regulon])):
        value = triplets[regulon][i]
        if value["predikat"] == "repressor":
            repressor = {'mainsnak': {'snaktype': 'value', 'property': "P3776", 'datavalue': {'value': {'entity-type': 'item', 'numeric-id': str(gene_ids[value["objekt"]][1:])}, 'type': 'wikibase-entityid'}}, 'type': 'statement', 'rank': 'normal'}
            claims.append(repressor.copy())
        elif value["predikat"] == "activator":
            activator_of = {'mainsnak': {'snaktype': 'value', 'property': "P3771", 'datavalue': {'value': {'entity-type': 'item', 'numeric-id': str(gene_ids[value["objekt"]][1:])}, 'type': 'wikibase-entityid'}}, 'type': 'statement', 'rank': 'normal'}
            claims.append(activator_of.copy())
        elif value["predikat"] == "dual":
            dual = {'mainsnak': {'snaktype': 'value', 'property': "P1322", 'datavalue': {'value': {'entity-type': 'item', 'numeric-id': str(gene_ids[value["objekt"]][1:])}, 'type': 'wikibase-entityid'}}, 'type': 'statement', 'rank': 'normal'}
            claims.append(dual.copy())

    # claim für found in taxon
    found_in_taxon = {'mainsnak': {'snaktype': 'value', 'property': "P703", 'datavalue': {'value': {'entity-type': 'item', 'numeric-id': 21102938}, 'type': 'wikibase-entityid'}}, 'type': 'statement', 'rank': 'normal'}
    claims.append(found_in_taxon)

    # claims für strand orientation
    if info["left_position"] != None:
        genomic_start = {'mainsnak': {'snaktype': 'value', 'property': "P644", 'datavalue': {'value': str(info["left_position"]), 'type': 'string'}, 'type': 'wikibase-entityid'}, 'type': 'statement', 'rank': 'normal'}
        claims.append(genomic_start)
    if info["right_position"] != None:
        genomic_end = {'mainsnak': {'snaktype': 'value', 'property': "P645", 'datavalue': {'value': str(info["right_position"]), 'type': 'string'}, 'type': 'wikibase-entityid'}, 'type': 'statement', 'rank': 'normal'}
        claims.append(genomic_end)
    if info["strand_orientation"] != None and info["left_position"] != None and info["right_position"] != None:
        switch = {"22809711" : "22809680", "22809680" : "22809711"}
        so = switch[str(info["strand_orientation"][1:])]
        strand_orientation = {'mainsnak': {'snaktype': 'value', 'property': "P2548", 'datavalue': {'value': {'entity-type': 'item', 'numeric-id': so}, 'type': 'wikibase-entityid'}}, 'type': 'statement', 'rank': 'normal'}
        claims.append(strand_orientation)

    # upload
    data = {'claims': claims}
    new_item.editEntity(data, summary='Adding multiple claims in one request')
    return new_item.getID()
```

Item wird erzeugt

Ergebnis


OxyS (Q127789260)
OxyS is a small RNA that plays a regulatory role in the oxidative stress response and has an antimutator effect.
In more languages
Configure
Language | Label | Description | Also known as
English | OxyS | OxyS is a small RNA that plays a regulatory role in the oxidative stress response and has an antimutator effect.
German | No label defined | No description defined
French | No label defined | No description defined
Bavarian | No label defined | No description defined
Statements
instance of
RNA
0 references
add reference
non-coding RNA
0 references
add reference
add value
inhibitor of
rpoS
0 references
add reference
fhlA
0 references
add reference
edit
Wikipedia (0 entries) edit
Wikibooks (0 entries) edit
Wikinews (0 entries) edit
Wikiquote (0 entries) edit
Wikisource (0 entries) edit
Wikiversity (0 entries) edit
Wikivoyage (0 entries) edit
Wiktionary (0 entries) edit
Multilingual sites (0 entries) edit
1 regulon_qids
{'OxyS': 'Q127789260',
'ChiX': 'Q127789264',
'SgrS': 'Q127789265',
'IstR-1': 'Q127789266',
'RseX': 'Q127789268',
'RydC': 'Q127789270',
'OhsC': 'Q127789272',
'SymR': 'Q127789274',
'MgrR': 'Q127789276',
'FnrS': 'Q127789277',
'ArrS': 'Q127789278',
'SdsN': 'Q127789281',
'AgrB': 'Q127789283',
'RalA': 'Q127789285',
'MicL-S': 'Q127789288',
'CpxQ': 'Q127789289',
'SdhX': 'Q127789292',
'C0293': 'Q127789294',