Diseño de la base de datos

Proyecto Single-Cell

Explicación del proyecto

Modulo Fuseki -Jena:

- Servicio apache
- Base de datos con una ontología.
- Consultas a traves de una API
- Consultas en SPARQL
- Repositorio común para proyectos de HCA y SCEA

Explicación del proyecto

Modulo API REST:

- Desarrollada en Python con Flask
- Permite acceder a los datos de Fuseki
- Hacer consultas de un modo más sencillo
- Documentación web con Swagger

Proyectos

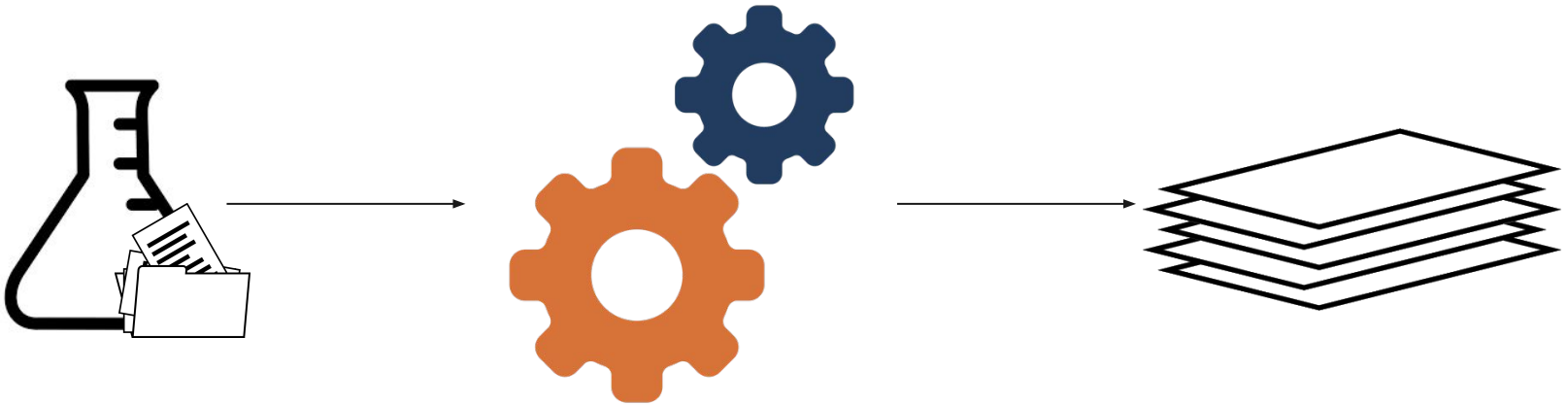
Metadatos:

- Enfermedad
- Tipo celular
- Órganos

Matriz de expresión:

- Tantas columnas como genes
- Tantas filas como células
- Matrices del orden de 100k x 20k

Percentiles y redes de co-expresión

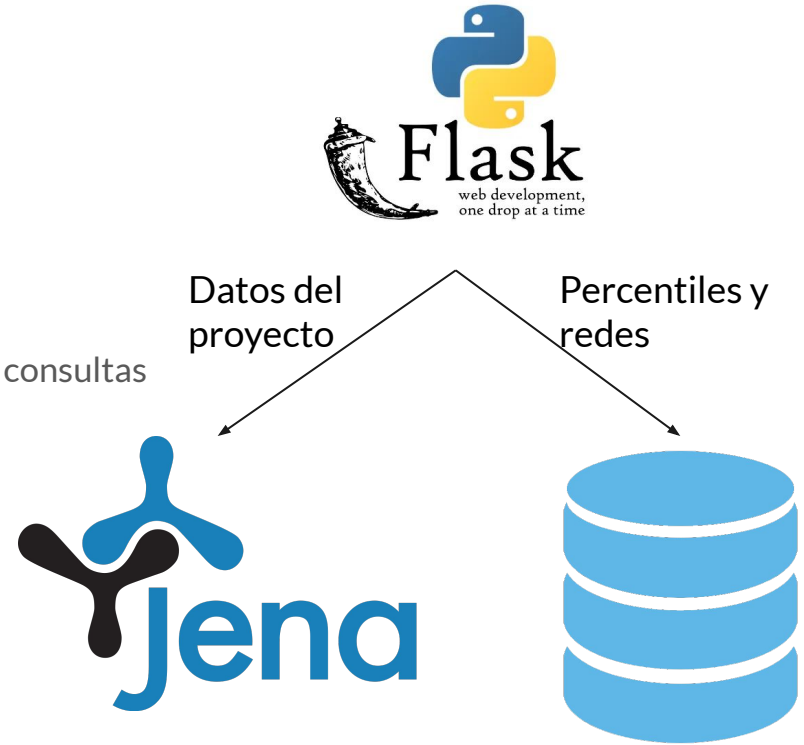


- Tendremos como mínimo tantos percentiles para un proyecto como genes haya en él.
- Las redes generarán como mínimo un Module Membership como genes haya en él.

Idea

Nuevo módulo de base de datos:

- Incorporará los datos de los percentiles
- Incorporará los datos de las redes de co-expresión
- La API-REST se comunicará con este módulo y le hará consultas
- Se podrá manipular la bd desde cualquier ordenador
- Automatizar la incorporación de percentiles y redes



¿Qué base de datos elegir? - Percentiles

Información a guardar de los percentiles:

- Nombre gen
- Percentil
- ID proyecto
- Num genes para calcular el percentil
- Num células para calcular el percentil
- Metadatos (tipo celular, organo, enfermedad...) dependiendo del proyecto

```
{
    'gen_name': 'ENSG00000288564',
    'percentile': 29.8354619721551,
    'project_ID': 'E-GEOD-36552',
    'gens': 26073,
    'cells': 123,
    'specie': 'homo sapiens',
    'cell_type': 'blastoderm cell',
    'organ': 'zygote'
}
```

¿Qué base de datos elegir? - Percentiles

Consultas sobre los percentiles:

- Consultas por nombre del gen
- Consultas por metadatos
- Consultas por proyecto

¿Qué base de datos elegir? - Redes de co-expresión

Información a guardar de las redes - MM:

- ID proyecto
- Metadatos (tipo celular, organo, enfermedad...) dependiendo del proyecto
- Corrección
- Iteración pseudocélulas
- Módulo
- Nombre del gen
- MM

```
{
    'ID_project': 'E-ENAD-20',
    'correccion': 'none',
    'iter_pseudocells': 0,
    'disease': 'melanoma',
    'organ': 'skin',
    'gen_name': 'AOC1',
    'MM': 0.23413091485896,
    'module': brown2
}
```

¿Qué base de datos elegir? - Redes de co-expresión

Módulo:

- Nombre (color)
- Notaciones
- Notaciones de fenotipos

```
{
    'name': 'brown2',
    'notations': [
        {
            'term': cytoplasm,
            'p-value': 0.0053
        }
    ],
    'phenotype notations': [
        {
            'term': Generalized hypotonia,
            'p-value': 0.00164
        }
    ],
}
```

Término:

- Nombre
- ID
- Fuente
- IC

```
{
    'name': 'cytoplasm',
    'ID': 'GO:0005737',
    'source': 'GO:CC',
    'IC': 1.14744902579554
}
```

Término de fenotipo :

- Nombre
- ID
- Fuente

```
{
    'name': 'Generalized hypotonia',
    'ID': 'HP:0001290',
    'source': 'HPO'
}
```

¿Qué base de datos elegir? - Redes de co-expresión

Consultas sobre los datos de las redes:

- Consultas por nombre del gen
- Consultas por metadatos
- Consultas por proyecto
- Consultas por términos
- ¿Consultas por p-valor?

¿Qué base de datos elegir?