Projeto Burials

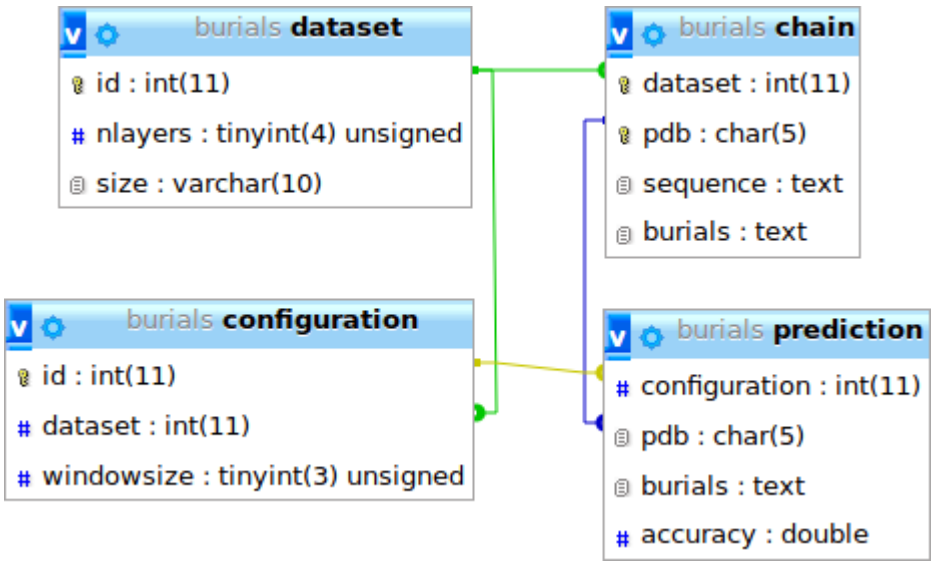
O objetivo desse projeto é construir uma aplicação web para facilitar a exploração de um banco de dados contendo informações sobre sequências de proteínas e os **enterramentos atômicos** (*burials*) de suas estruturas.

Enterramentos atômicos, definidos como a distância de cada átomo da proteína ao centro geométrico da estrutura, são uma medida estrutural definida pelo nosso grupo de pesquisa (VAN DER LINDEN et al., 2013, FERREIRA et al., 2016) como um intermediário informacional entre sequência e estrutura, com o objetivo de ser usado em algoritmos de predição de estruturas de proteínas a partir das sequências.

1. Instalando o Banco de dados

1. Criar um novo banco de dados no MySQL
2. Nesse banco, executar o script `create-burials-database.sql` (para criar as tabelas)
3. Executar o script `insert-burial-data.sql` (para popular o banco – esse passo pode demorar)

2. Estrutura do Banco de dados



A unidade básica do banco de dados utilizado é a cadeia (*chain*), que representa uma estrutura de proteína. Cada *chain* pertence a um único *dataset* e está associada a várias predições de

enterramentos (*predictions*). Cada uma dessas predições resulta em uma sequência de enterramentos (campo *burials*) e um nível de acurácia (campo *accuracy*) e está associada a 2 parâmentros: um tamanho da janela, que pertence a uma configuração específica (campo *windowsize* da tabela *configuration*), e uma divisão em número de camadas (campo *nlayers* da tabela *dataset*, associada à respectiva *configuration*).

3. Consultas que devem ser possíveis na aplicação web

3.1. Buscar por todas as predições de acordo com um conjunto de parâmetros

Parâmetros:

1. Número de camadas (entre 3 e 5)
2. Classe de tamanho (0, 80, 50_100, 120, 160, 240 ou 1000)
3. Tamanho da janela (entre 1 e 5)

Segue um exemplo de consulta SQL para a busca com **5** camadas, classe de tamanho **240** e tamanho da janela **4**:

```
SELECT p.pdb, ch.sequence, ch.burials, p.burials AS prediction, p.accur
acy
FROM prediction AS p
JOIN configuration AS cf ON p.configuration = cf.id
JOIN dataset as ds ON cf.dataset = ds.id
JOIN chain as ch ON p.pdb = ch.pdb
WHERE
ch.dataset = ds.id
AND ds.nlayers=5
AND ds.size='240'
AND cf.windowsize=4
```

3.2. Buscar a acurácia média por tamanho da janela

Parâmetros:

1. Número de camadas (entre 3 e 5)
2. Classe de tamanho (0, 80, 50_100, 120, 160, 240 ou 1000)

Exemplo para **4** camadas e classe de tamanho **0**:

```
SELECT cf.windowsize, AVG(p.accuracy) AS avgacc
FROM prediction AS p
JOIN configuration AS cf ON p.configuration = cf.id
JOIN dataset as ds ON cf.dataset = ds.id
JOIN chain as ch ON p.pdb = ch.pdb
WHERE
ch.dataset = ds.id
AND ds.nlayers=4
AND ds.size='0'
GROUP BY (cf.windowsize)
```

3.3. Buscar a acurácia média por número de camadas

Parâmetros:

1. Classe de tamanho (0, 80, 50_100, 120, 160, 240 ou 1000)
2. Tamanho da janela (entre 1 e 5)

Exemplo para classe de tamanho **240** e tamanho da janela **5**:

```
SELECT ds.nlayers, AVG(p.accuracy) AS avgacc
FROM prediction AS p
JOIN configuration AS cf ON p.configuration = cf.id
JOIN dataset as ds ON cf.dataset = ds.id
JOIN chain as ch ON p.pdb = ch.pdb
WHERE
ch.dataset = ds.id
AND cf.windowsize=5
AND ds.size='240'
```

```
GROUP BY (ds.nlayers)
```

4. Referências

VAN DER LINDEN, MARX GOMES ; FERREIRA, DIOGO CÉSAR ; DE OLIVEIRA, LEANDRO CRISTANTE ; ONUCHIC, JOSÉ N. ; DE ARAÚJO, ANTÔNIO F. PEREIRA . Ab initio protein folding simulations using atomic burials as informational intermediates between sequence and structure. Proteins (Print) , v. 82, 2013.

FERREIRA, DIOGO C. ; VAN DER LINDEN, MARX G. ; DE OLIVEIRA, LEANDRO C. ; ONUCHIC, JOSÉ N. ; PEREIRA DE ARAÚJO, ANTÔNIO F. . Information and redundancy in the burial folding code of globular proteins within a wide range of shapes and sizes. Proteins (Print) , v. 84, p. 515-531, 2016.