PROJETO EM BIOINFORMÁTICA

Projeto 23

Towards a public repository of designed proteins

Karolina Lopes Barbosa (PG55129)

Introdução

PROTEIN DESIGN

O design de proteínas permite criar proteínas sintéticas ou alterar proteínas já existentes para que tenham novas funcionalidades.

RELEVÂNCIA

São úteis por exemplo, no desenvolvimento de fármacos, vacinas e outras aplicações biotecnológicas.

Diversas técnicas e metodologias são utilizadas para projetar essas proteínas, o que gera alta variedade de dados experimentais e computacionais.

O que se torna em um desafio, pois esses dados permanecem descentralizados.

Objetivo

Esse desafio evidencia a necessidade da criação de uma base de dados centralizada e de acesso público, capaz de reunir informações para proteínas sintéticas.

- Na fase inicial, o objetivo é mapear os esforços em andamento e os repositórios já existentes relacionados ao design de proteínas, a fim de compreender melhor o panorama atual da área.
- A próxima etapa consiste na implementação da base de dados, e no desenvolvimento de uma aplicação que permitirá sua interação via interface web.

State of the Art

Na pesquisa por repositórios, foram identificadas algumas bases de dados dedicadas ao armazenamento de proteínas artificiais:

- The Protein Design Archive

  Parece ser um repositório voltado prioritariamente para o armazenamento de estruturas determinadas experimentalmente.

- Protαβank

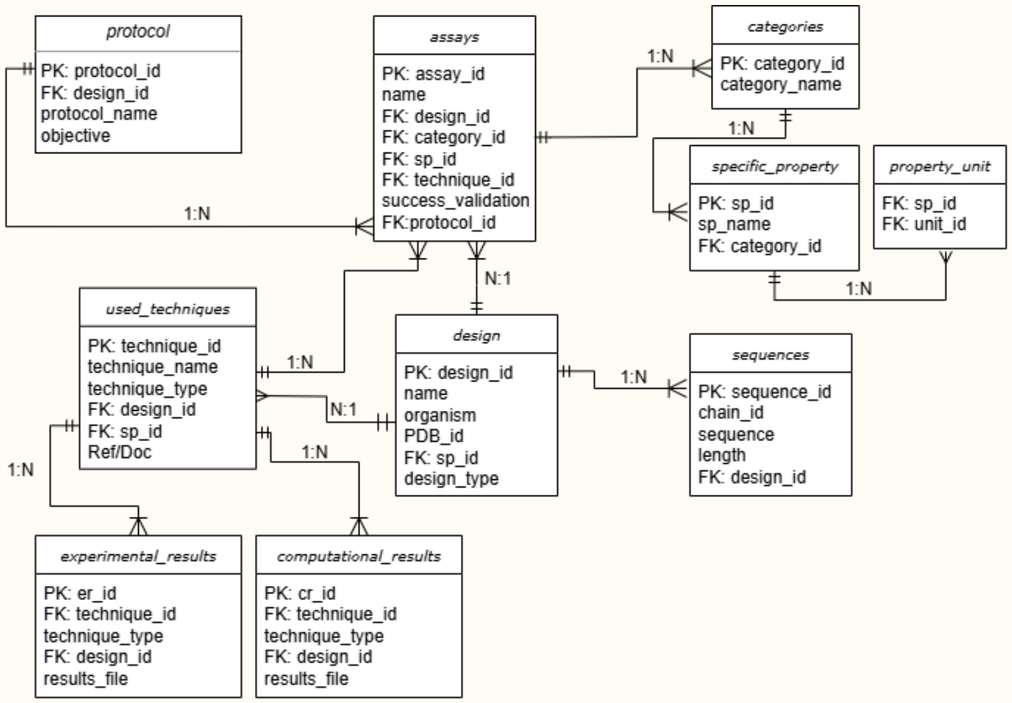
  Consideramos mais completo em relação ao objetivo do projeto, essa base integra dados experimentais e computacionais. Apesar disso, pertence a uma empresa privada e parece não apresentar atualizações desde 2022.

Métodos

Considerando a existência do ProtaBank, sua estrutura foi utilizada como referência para o desenvolvimento da base de dados proposta.

ESQUEMA DA BASE DE DADOS

Um esquema inicial da base de dados foi elaborado para guiar sua construção, identificando as informações relevantes e estabelecendo os relacionamentos entre as entidades.



Métodos

FERRAMENTAS UTILIZADAS

DATABASE

BACKEND+FRONTEND

+ django

Métodos

CONSTRUÇÃO DA BASE DE DADOS

- Durante a construção da base de dados, foram realizados testes para verificar os relacionamentos entre as tabelas, validar a integridade dos dados e assegurar o correto funcionamento das dependências de chave estrangeira.

DESAFIO: POSSIBILITAR UMA INGESTÃO DE DADOS USER-FRIENDLY

- Como o design de uma única proteína pode gerar grandes volumes de dados, o processo de inserção pode se tornar exaustivo para o usuário (uma limitação observada no ProtaBank). Assim, um dos desafios desta base de dados foi desenvolver uma solução que torne a inserção de informações mais prática e eficiente.

Métodos

DESENVOLVIMENTO DE INTERFACE WEB

- Com o objetivo de simplificar o processo de inserção e consulta dos dados das proteínas, foi criada uma interface web.

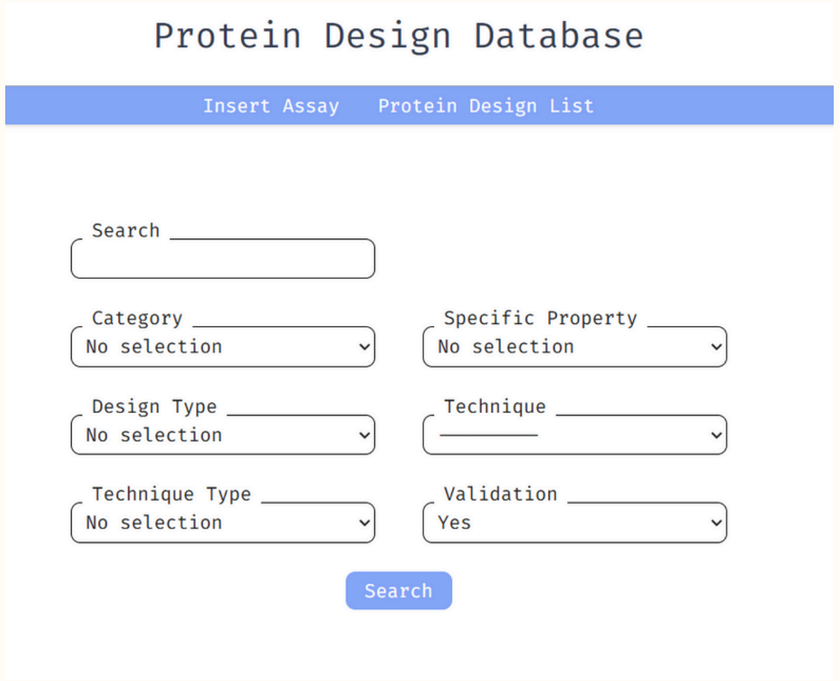
- O framework Django foi utilizado para criar uma aplicação e integrar a base de dados MySQL a 4 páginas web desenvolvidas.

Resultados

PROTEIN DESIGN DATABASE

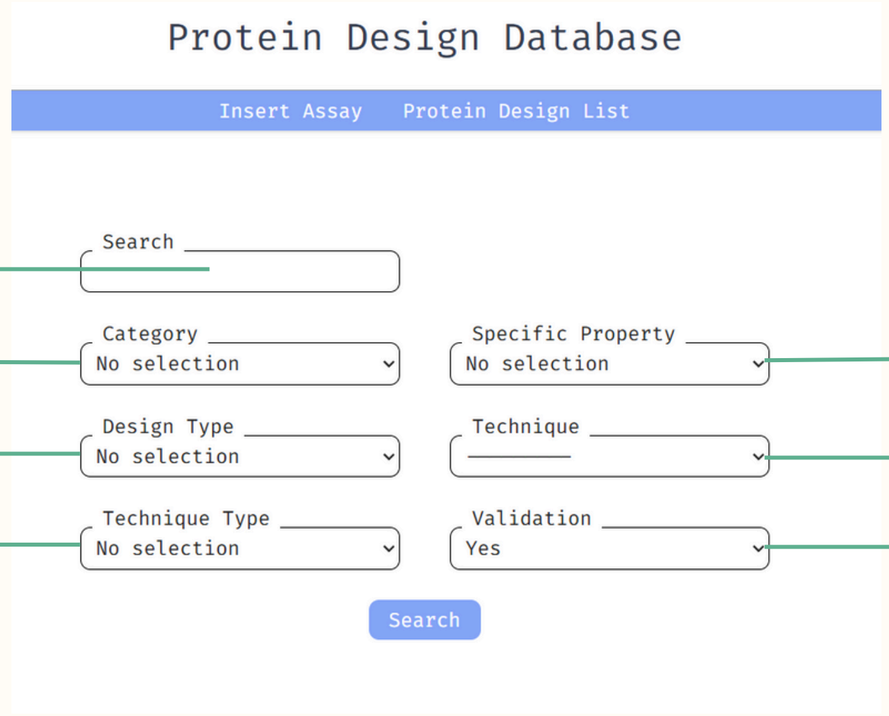
FILTROS DE BUSCA

- Interface de busca com filtros para localizar proteínas previamente inseridas na base de dados.



Resultados

PROTEIN DESIGN DATABASE: FILTROS DE BUSCA



Nome da proteína; Organismo; PDB ID.

Categoria do Design

Protein Engineering ou De Novo Design

Experimental ou Computacional

Propriedade Específica do Design

Nome da Técnica

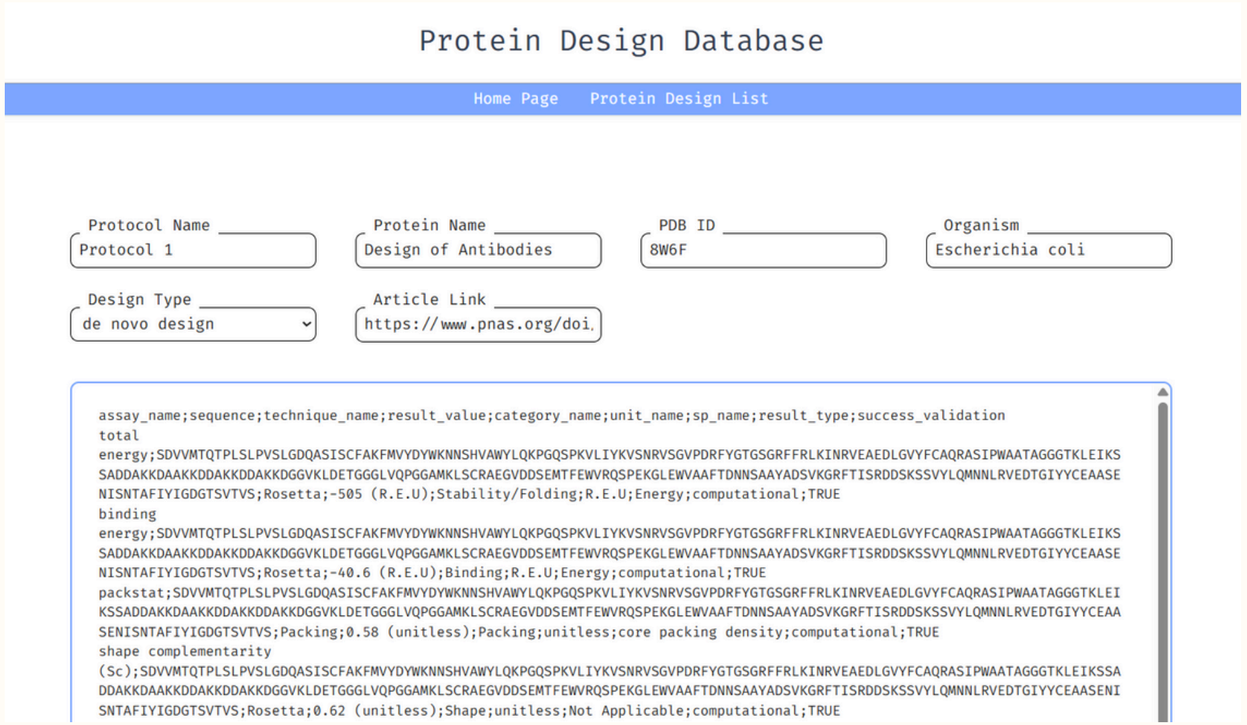
Sim ou Não para Validação de Ensaios

Resultados

PROTEIN DESIGN DATABASE: INSERIR PROTEÍNA

Interface para a inserção dos dados para protein design:

- Formulário para inserção de informações gerais sobre o design da proteína.
- Formulário de inserção em lote (bulk data form em .csv) para inserção de informações experimentais e computacionais associadas a proteína.



Resultados

PROTEIN DESIGN DATABASE: INSERIR PROTEÍNA

- Cada design pode ter uma grande variabilidade de dados

BULK DATA FORM

9 colunas fixas que permitem a inserção de diversos dados relacionados a cada uma delas.

```
assay_name;sequence;technique_name;result_value;category_name;unit_name;sp_name;result_type;success_validation
total
energy;SDVVMTQTPLSLPVSLGDQASISCFAKFMVYDYWKNNSHVAWYLQKPGQSPKVLIYKVSNRVSGVPDRFYGTGSGRFFRLKINRVEAEDLGVYFCAQRASIPWAATAGGGTKLEIKS
SADDAKKDAAKKDDAKKDDAKKDGGVKLDETGGGLVQPGGAMKLSCRAEGVDDSEMTFEWVRQSPEKGLEWVAAFTDNNSAAYADSVKGRFTISRDDSKSSVYLQMNNLRVEDTGIYYCEAASE
NISNTAFIYIGDGTSVTVS;Rosetta;-505 (R.E.U);Stability/Folding;R.E.U;Energy;computational;TRUE
binding
energy;SDVVMTQTPLSLPVSLGDQASISCFAKFMVYDYWKNNSHVAWYLQKPGQSPKVLIYKVSNRVSGVPDRFYGTGSGRFFRLKINRVEAEDLGVYFCAQRASIPWAATAGGGTKLEIKS
SADDAKKDAAKKDDAKKDDAKKDGGVKLDETGGGLVQPGGAMKLSCRAEGVDDSEMTFEWVRQSPEKGLEWVAAFTDNNSAAYADSVKGRFTISRDDSKSSVYLQMNNLRVEDTGIYYCEAASE
NISNTAFIYIGDGTSVTVS;Rosetta;-40.6 (R.E.U);Binding;R.E.U;Energy;computational;TRUE
packstat;SDVVMTQTPLSLPVSLGDQASISCFAKFMVYDYWKNNSHVAWYLQKPGQSPKVLIYKVSNRVSGVPDRFYGTGSGRFFRLKINRVEAEDLGVYFCAQRASIPWAATAGGGTKLEI
KSSADDAKKDAAKKDDAKKDDAKKDGGVKLDETGGGLVQPGGAMKLSCRAEGVDDSEMTFEWVRQSPEKGLEWVAAFTDNNSAAYADSVKGRFTISRDDSKSSVYLQMNNLRVEDTGIYYCEAA
SENISNTAFIYIGDGTSVTVS;Packing;0.58 (unitless);Packing;unitless;core packing density;computational;TRUE
shape complementarity
(Sc);SDVVMTQTPLSLPVSLGDQASISCFAKFMVYDYWKNNSHVAWYLQKPGQSPKVLIYKVSNRVSGVPDRFYGTGSGRFFRLKINRVEAEDLGVYFCAQRASIPWAATAGGGTKLEIKSSA
DDAKKDAAKKDDAKKDDAKKDGGVKLDETGGGLVQPGGAMKLSCRAEGVDDSEMTFEWVRQSPEKGLEWVAAFTDNNSAAYADSVKGRFTISRDDSKSSVYLQMNNLRVEDTGIYYCEAASENI
SNTAFIYIGDGTSVTVS;Rosetta;0.62 (unitless);Shape;unitless;Not Applicable;computational;TRUE
```

Resultados

PROTEIN DESIGN DATABASE: INFORMAÇÕES DA PROTEÍNA

Protein Name: Design of antibodies

PDB ID: 5NBI

Organism: Escherichia coli

Design Type: De Novo Design

Article Link: https://www.pnas.org/doi/10.1073/pnas.1707171

Used Techniques

1: Rosetta

2: Packing

3: Fluorescence Intensity

Categories & Specific Properties

#	Category	Specific Properties
1	Binding	• Energy
2	Stability/Folding	• Energy
3	Packing	• core packing density
4	Shape	• Not Applicable • shape
5	Expression	• Not Applicable
6	accessibility to solvent	• solvent accessible surface area • Not Applicable

Sequences

Sequence 1:
Length: 260 aa
SDVVMTQTPLSLPVSLGDQASISCFAKFMVYDYWKNNSHVAWYLQKPGQSPKVLIYKVSNRVSGVPDRFYGTGSGRFFRLKINRVEAEDLGVYFCAQRASIPWAATAGGGTKLEIKSSADDAKKDAAKKDDAKKDDAKKDGG\

Sequence 2:
Length: 260 aa
SDVVMTQTPLSLPVSLGDQASISCFAKFMVYDYWKNNSHVAWYLQKPGQSPKVLIYKVSNRVSGVPDRFYGTGSGRFFRLKINRVEAEDLGVYFCAQRASIPWAATAGGGTKLEIKSSADDAKKDAAKKDDAKKDDAKKDGG\

Sequence 3:
Length: 260 aa
SDVVMTQTPLSLPVSLGDQASISCFAKFMVYDYWKNNSHVAWYLQKPGQSPKVLIYKVSNRVSGVPDRFYGTGSGRFFRLKINRVEAEDLGVYFCAQRASIPWAATAGGGTKLEIKSSADDAKKDAAKKDDAKKDDAKKDGG\

Sequence 4:
Length: 260 aa
SDVVMTQTPLSLPVSLGDQASISCFAKFMVYDYWKNNSHVAWYLQKPGQSPKVLIYKVSNRVSGVPDRFYGTGSGRFFRLKINRVEAEDLGVYFCAQRASIPWAATAGGGTKLEIKSSADDAKKDAAKKDDAKKDDAKKDGG\

Assays

Assay 1: total energy
Validation: True

Assay 2: binding energy
Validation: True

Assay 3: packstat
Validation: True

Assay 4: shape complementarity (Sc)

Computational Results

Computational Result 1:
Value: -505 (R.E.U)

Computational Result 2:
Value: -40.6 (R.E.U)

Computational Result 3:
Value: 0.58 (unitless)

Computational Result 4:
Value: 0.62 (unitless)

Experimental Results

Experimental Result 1:
Value: 11.3%

Experimental Result 2:
Value: 2.6%

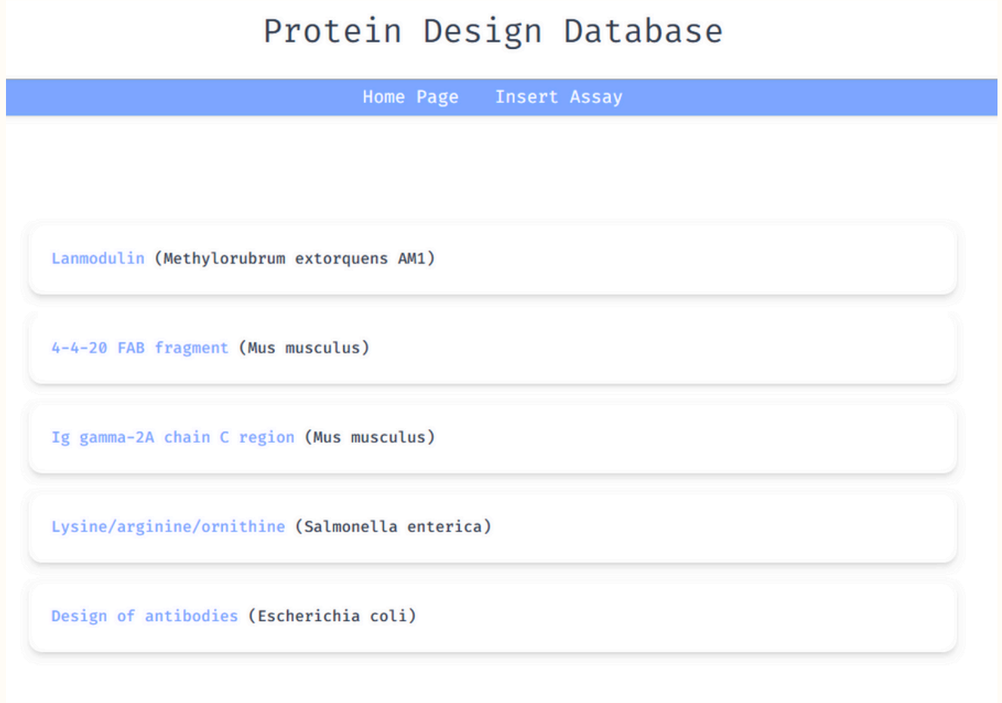
Experimental Result 3:
Value: 3.2%

Experimental Result 4:
Value: 0.2%

Resultados

PROTEIN DESIGN: LISTA DE PROTEÍNAS

- Página dedicada à visualização de todas as proteínas cadastradas na base de dados.



Discussão e Conclusão

- Os exemplos de dados inseridos na base de dados foram obtidos a partir do ProtaBank.

- A utilização da interface web possibilitou uma visualização mais clara dos dados, contribuindo para a identificação de futuras melhorias nos relacionamentos entre as tabelas da base de dados, para aprimorar a organização e a eficiência das consultas.

- Algumas validações básicas foram implementadas para a inserção de dados, porém ainda são necessárias diversas outras validações para assegurar que as informações sejam inseridas de forma consistente e correta.

- A página de busca pode ser otimizada para buscar também sequências das proteínas.

Discussão e Conclusão

- Outras informações relevantes para uma base de dados voltada para design de proteínas poderiam ser adicionadas, de forma a ampliar a variedade de dados armazenados.

- Apesar do bulk data form ter facilitado a inserção de dados, ainda é complexo inserir a alta variedade de dados que um design pode gerar.
  - ↳ Ainda dado a variabilidade e a natureza flexível dos dados, consideramos a hipótese de que a utilização de um modelo NoSQL poderia proporcionar uma implementação mais eficiente para a base de dados.

Obrigada!