Zámer tímového projektu

**Študijný program:** Softwarové a datové inženýrství

**Typ projektu:** Softwarový projekt (NPRG069)

**Študenti:** Katarína Bucková, Richard Fedák, Samuel Karaš, Milan Truchan

**Vedúci:** doc. RNDr. David Hoksza, Ph.D.

**Konzultanti:** Mgr. Marian Novotný, Ph.D., Bc. Lukáš Polák, Mgr. Petr Škoda, Ph.D.

**Názov a téma projektu:** Rozšírenie projektu Prankweb (TODO – iný názov)

**Prehľad**

Cieľom práce je rozšírenie existujúceho projektu Prankweb, ktorý umožňuje, okrem iného, na základe štruktúr proteínov predikovať a vizualizovať potenciálne väzobné miesta ligandov. Rozšírenie spočíva v pridaní nového spôsobu nachádzania väzobných miest pomocou

1. pLM (protein Language Model), ktorý nám vráti embeddings – feature vektory aminokyselín. Každý embedding bude ako vstup pre neurónovú sieť, ktorá bude binárne klasifikovať či daná aminokyselina patrí alebo nepatrí do rezídua/väzobného miesta

(**NOTES** datasety su k dispozicii, pLM-ESM2, neuronovú siet – natrenovat+vytvorit)

1. AHoJ-DB obsahujúcej predpočítané štruktúry pre danú sekvenciu. Ku každej štruktúre databáza poskytuje dva typy väzobných miest pre ligandy:
   1. APO (ligand free) - miesto s nenaviazaným ligandom
   2. HOLO (ligand bound) - miesto s naviazaným ligandom

Protein može mat viac vazobnych miest,

Api calls – mame seq a data z ahojdb – z toho ziskame všetkz možne proteinove strukturz a ich pokety, a ligandy na ktorych su poziciach

1. MMseqs2 obsahujúcej homologické sekvencie iných živočíšnych druhov, z ktorej je možné získať podobné sekvencie-štruktury (tie sekvencie nemusia byt uplne nepokryju celu cekvenciu) a väzobné miesta pre danú sekvenciu

Ďalším rozšírením je prevod sekvencie na 3D štruktúru(TODO AlphaFold?).

Projekt je rozdelený na štyri časti: Frontend, Backend, Dátová časť, Integrácia pLM.

**Frontend**

Cieľom je vizualizovať štruktúry proteínov a ich väzobné miesta (predikované zo sekvencií) (TODO aj s ligandami?). Súčasťou frontendu je aj vrstevnatá vizualizácia viacerých štruktúr/(TODO sekvencií?) s predikovanými väzobnými miestami. Užívateľ si bude môcť zvoliť štruktúry/(TODO sekvencie?), ktoré sa mu zobrazia.

**Backend**

Rozšírenie API pre funkcionality potrebné na fungovanie frontendu, pridanie dockerových kontajnerov a ich prepojenie s existujúcou architektúrou.

**Dátová časť**

Práca s AHoJ-DB a MMseqs2, vytvorenie šablón dotazov.

**Integrácia pLM**

Využitie existujúceho pLM na predikciu rezíduí (väzobných miest) na sekvencii užívateľom zadaného proteínu.

**Približný priebeh**

Odhadovaná dĺžka projektu je štandardných 9 mesiacov. Časti projektu sú nasledujúce:

1. Oboznámenie sa s doménou a štruktúrou projektu
2. Definícia funkčných požiadaviek
3. Tvorba špecifikácie projektu
4. Implementácia rozšírenia
5. Návrh a vývoj prípadných dodatočných funkcionalít
6. Testovanie
7. Tvorba dokumentácie