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:** Plankweb (rozšírenie projektu Prankweb)

**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 existujúceho **pLM (protein Language Model),** ktorý nám vráti embeddings – feature vektory aminokyselín. Každý embedding bude vstupom pre klasifikačnú neurónovú sieť (tú je nutné vytvoriť), ktorá rozhodne, či daná aminokyselina patrí alebo nepatrí do väzobného miesta.

Okrem toho bude môcť užívateľ využívať na skúmanie proteínov a ich (potenciálnych) väzobných miest dáta z:

1. **AHoJ-DB**, ktorá obsahuje predpočítané proteínové štruktúry. Pomocou AHoJ-DB vieme nájsť štruktúry pre danú sekvenciu proteínu. 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
2. **MMseqs2**, ktorá obsahuje homologické sekvencie iných živočíšnych druhov. Vďaka MMseqs2 dokážeme nájsť pre danú sekvenciu podobné sekvencie (tie nemusia byť úplné, môžu obsahovať „diery“) spolu s ich väzobnými miestami.

Ďalej rozšírenie zahŕňa aj prevod sekvencie na 3D štruktúru.

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

**Frontend**

Cieľom je vizualizovať štruktúry proteínov a ich väzobné miesta (predikované zo sekvencií). Súčasťou frontendu je aj vrstevnatá vizualizácia viacerých štruktúr s predikovanými väzobnými miestami (užívateľ si bude môcť vybrať, aby sa mu zobrazila vizualizácia viacerých proteínových štruktúr súčasne).

**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.

**AI**

Využitie existujúceho pLM a vytvorenie klasifikačnej neurónovej siete pre predikciu väzobných miest.

**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