

Baltymų struktūrų modeliavimas trimatėje erdvėje

Paulius Milmantas

Darbo vadovas: doc. dr. Linas Petkevičius

Vilniaus Universitetas

Matematikos ir informatikos fakultetas

Magistro baigiamojo darbo gynimas

Tiriama sritis (1)

Kas yra tiriama?

Darbe yra tiriami galimi baltymų struktūrų modeliavimo algoritmo ColabFold pagerinimo būdai. Vienas iš būdų - pradinės MSA struktūros sukūrimo metodo tobulinimas.

Kaip veikia modeliavimo algoritmai?

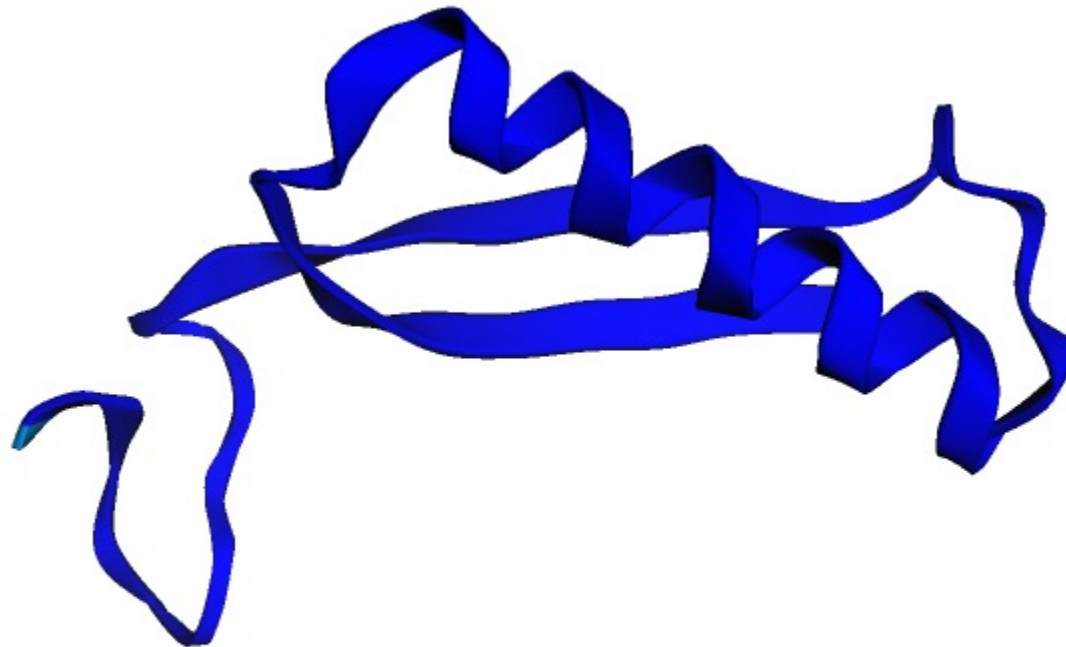
Baltymus sudaro aminorūgščių sekos. Trimatės erdvės baltymų modeliavimo algoritmai priima šias sekas ir grąžina atomų koordinates 3D erdvėje.

Kodėl tai yra svarbu?

Trimatis baltymų modeliavimas yra reikalingas, norint sužinoti baltymo struktūrą ir jo funkcinėmis savybėmis. Nuo to, kaip susilankstys baltymas trimatėje erdvėje, priklauso jo funkcionalumas.

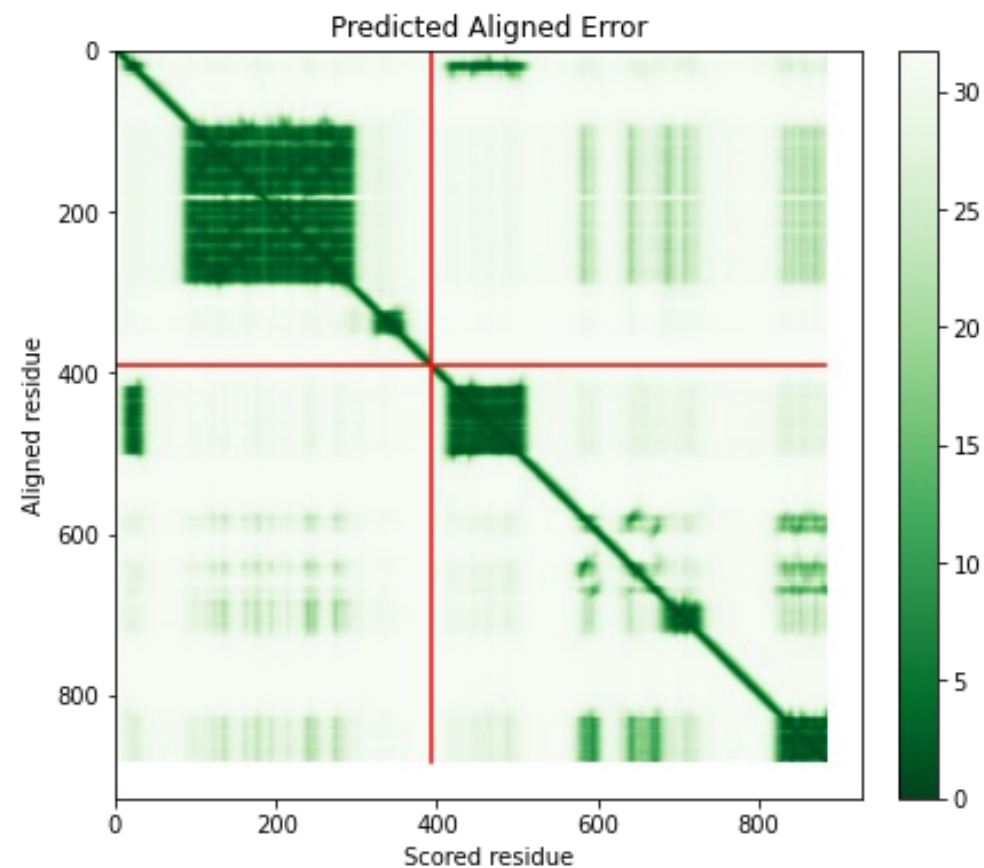
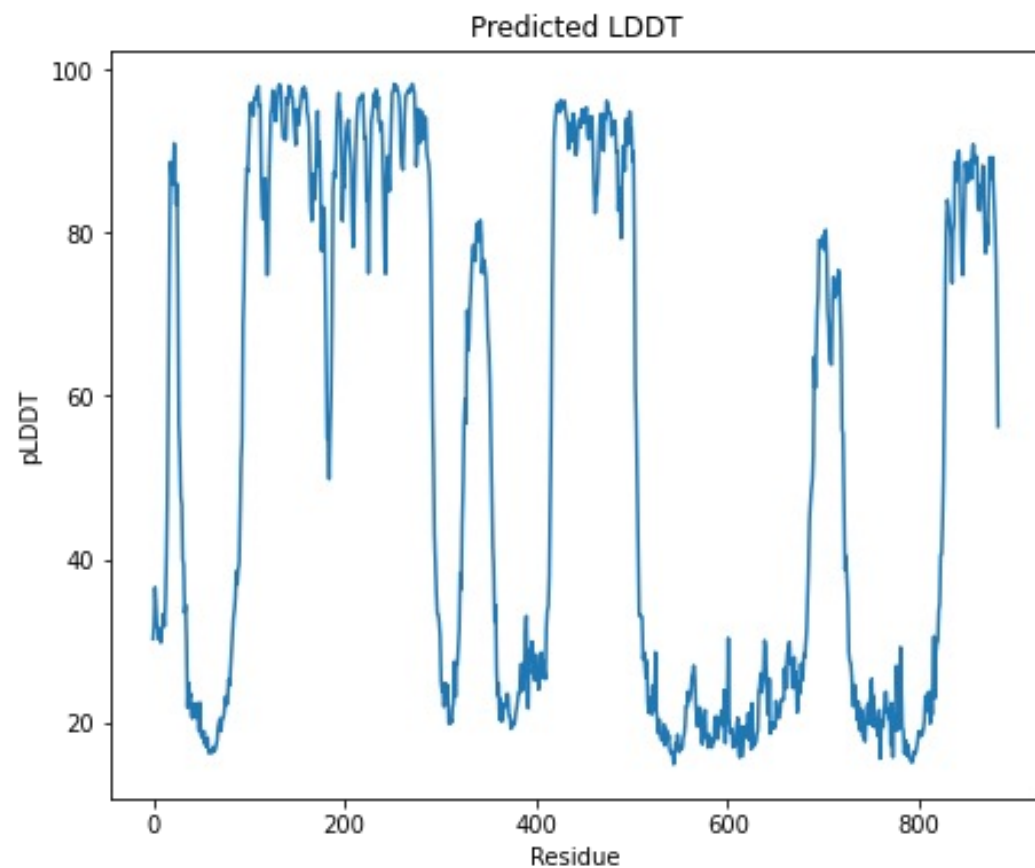
Tiriama sritis (2)

PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASK



pLDDT: ■ Very low (<50) ■ Low (60) ■ OK (70) ■ Confident (80) ■ Very high (>90)

Tiriama sritis (3)



Pav. 2 Klaidų tikimybių metrikos

Tikslas ir uždaviniai

Iškeltas pagrindinis tikslas - pasiūlyti ir validuoti matematinius ir infrastruktūrinius sprendimus, siekiant pagerinti baltymų modeliavimo algoritmus, patobulinant daugybinio sekų palyginio (MSA) sudarymą.

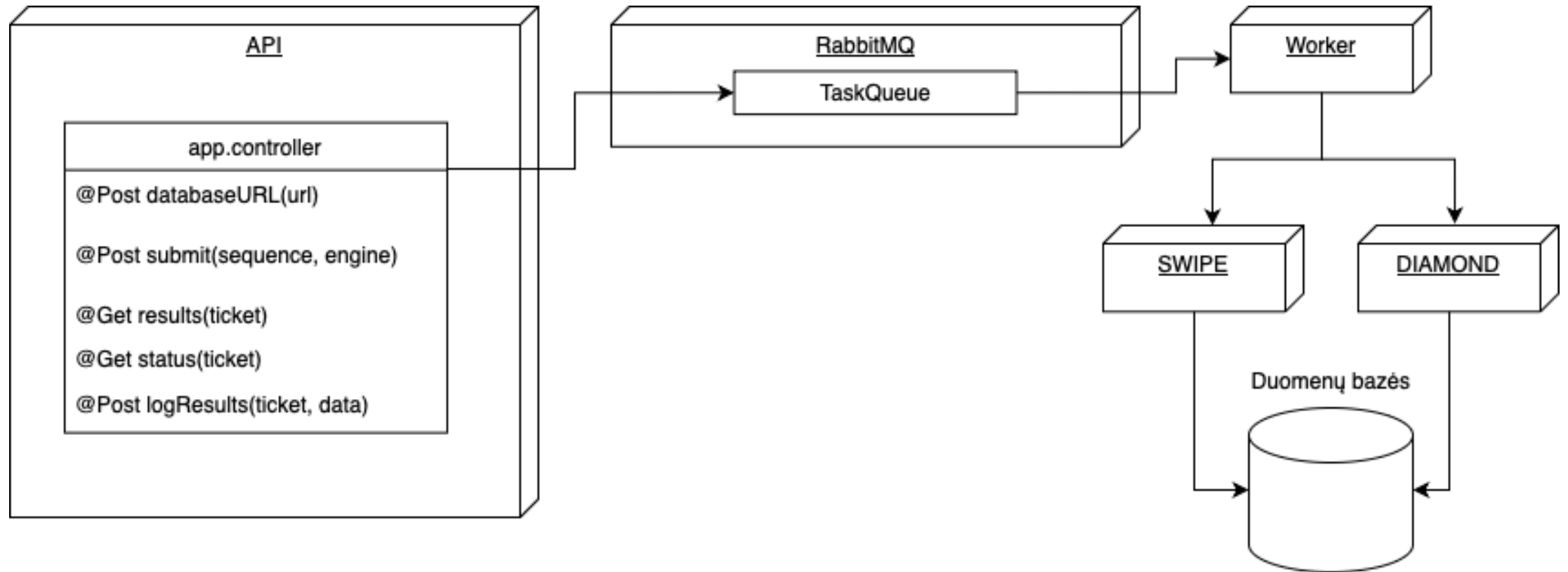
Darbe iškelti uždaviniai:

- Išanalizuoti ColabFold veikimą.
- Pateikti pasiūlymus, kaip galima patobulinti pradinių duomenų rinkinio sudarymo ColabFold algoritmą.
- Įgyvendinti pateiktus pasiūlymus.
- Palyginti įgyvendintų pasiūlymų grąžinamų rezultatų pLDDT metrikas su dabartiniu ColabFold sprendimu

Problematika

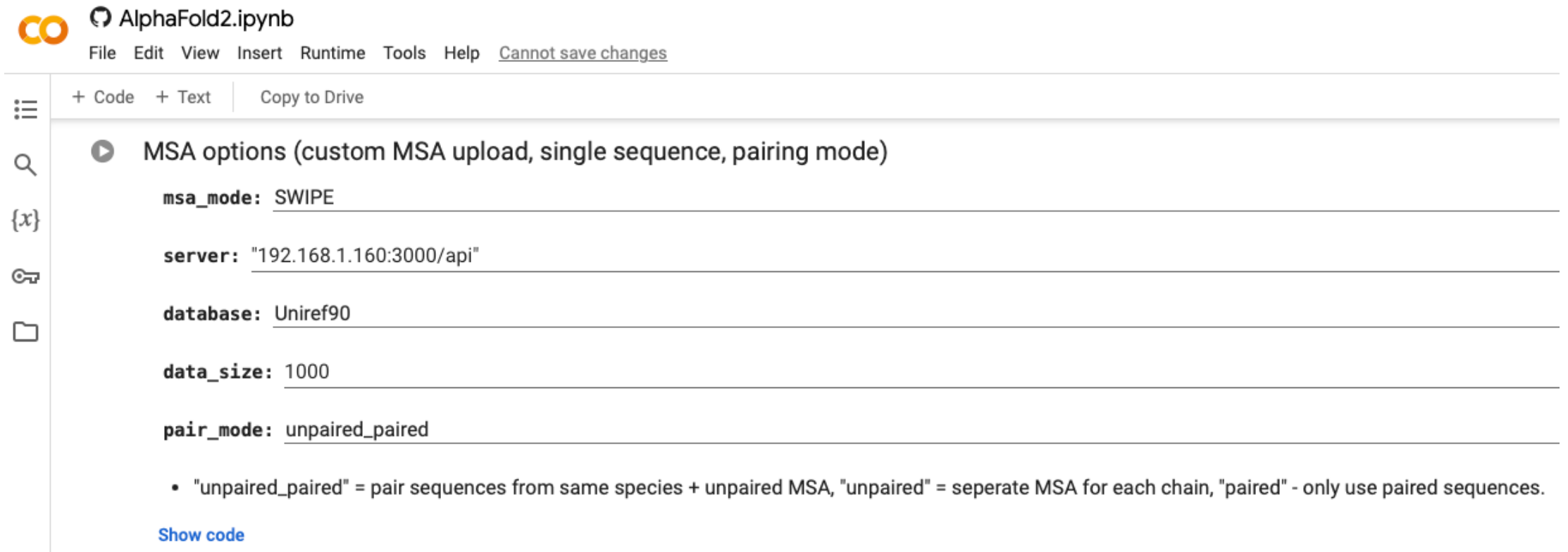
- ColabFold neleidžia keisti sekų paieškos algoritmo ar lengvai pridėti savo duomenų bazės, programos veikimo metu.
- ColabFold naudoja tik vieną sekų lyginimo algoritmą – MMSeq2.
- ColabFold neleidžia pakeisti sekų paieškos serverio adresą.

Pateiktas pasiūlymas (1)



Pav. 3 Pateikto sprendimo schema

Pateiktas sprendimas (2)



The screenshot shows a Google Colab notebook titled "AlphaFold2.ipynb". The interface includes a top menu bar with "File", "Edit", "View", "Insert", "Runtime", "Tools", "Help", and a status message "Cannot save changes". Below the menu is a toolbar with "+ Code", "+ Text", and "Copy to Drive". The notebook content is a code cell titled "MSA options (custom MSA upload, single sequence, pairing mode)". It contains several configuration lines: "msa_mode: SWIPE", "server: '192.168.1.160:3000/api'", "database: Uniref90", "data_size: 1000", and "pair_mode: unpaired_paired". A bullet point explains the "pair_mode" options: "unpaired_paired" = pair sequences from same species + unpaired MSA, "unpaired" = separate MSA for each chain, "paired" = only use paired sequences. A "Show code" link is at the bottom left of the cell.

AlphaFold2.ipynb
File Edit View Insert Runtime Tools Help Cannot save changes

+ Code + Text | Copy to Drive

▶ MSA options (custom MSA upload, single sequence, pairing mode)

msa_mode: SWIPE

server: "192.168.1.160:3000/api"

database: Uniref90

data_size: 1000

pair_mode: unpaired_paired

- "unpaired_paired" = pair sequences from same species + unpaired MSA, "unpaired" = separate MSA for each chain, "paired" - only use paired sequences.

[Show code](#)

Pav. 4 Patobulintas ColabFold

Rezultatai (1)

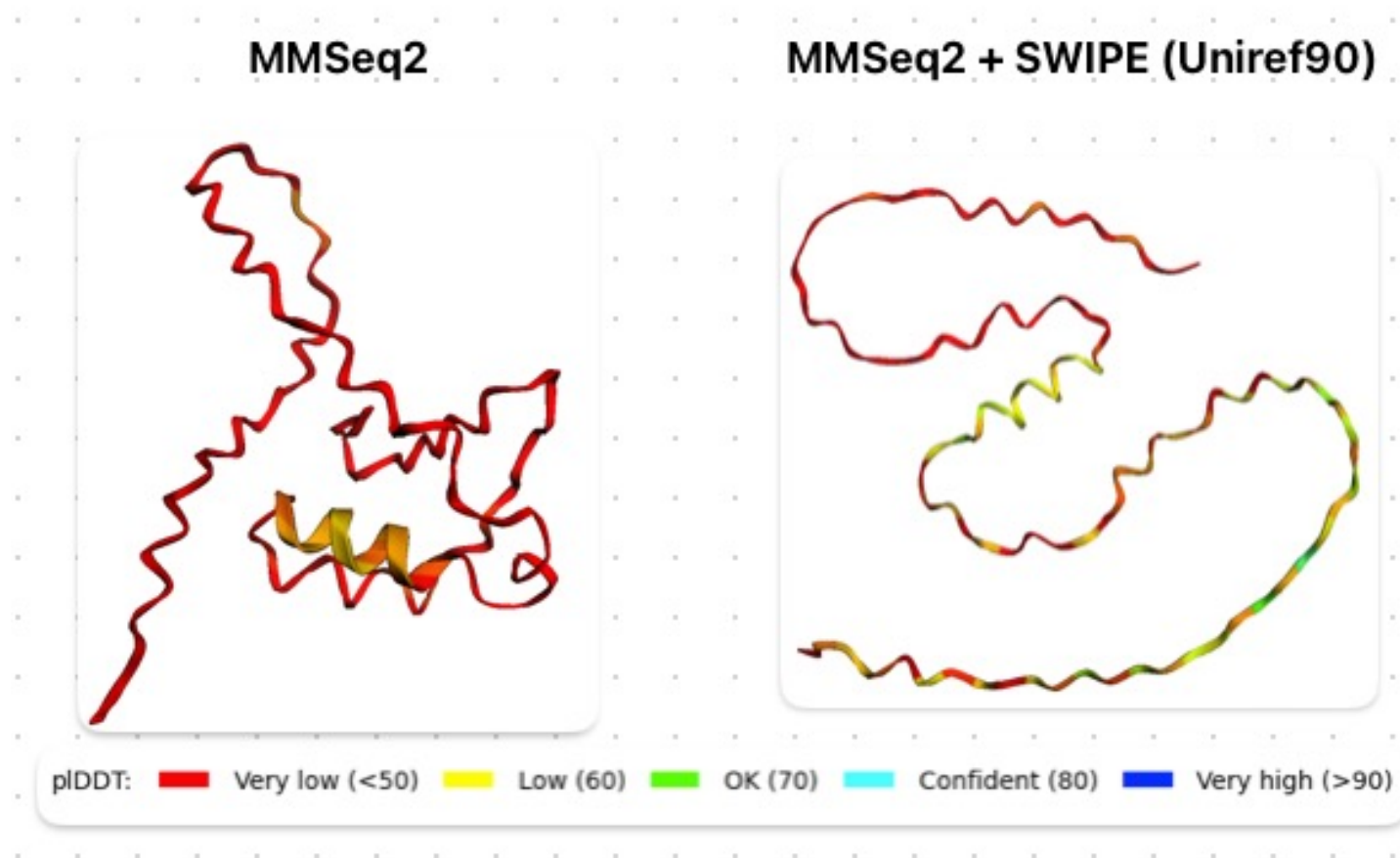
Algoritmas	Didžiausia pLDDT reikšmė
MMSeq2	97,4
MMSeq2	97,4
MMSeq2	97,4
MMSeq2 + DIAMOND	97,4
MMSeq2 + DIAMOND	97,4
MMSeq2 + DIAMOND	97,4
MMSeq2 + SWIPE	97,4
MMSeq2 + SWIPE	97,4
MMSeq2 + SWIPE	97,4

Lentelė 1. 1-as bandymas

Algoritmas	Didžiausia pLDDT reikšmė
MMSeq2	43,3
MMSeq2	43,3
MMSeq2	43,3
MMSeq2 + DIAMOND	43,3
MMSeq2 + DIAMOND	43,3
MMSeq2 + DIAMOND	43,3
MMSeq2 + SWIPE	56,1
MMSeq2 + SWIPE	56,1
MMSeq2 + SWIPE	56,1

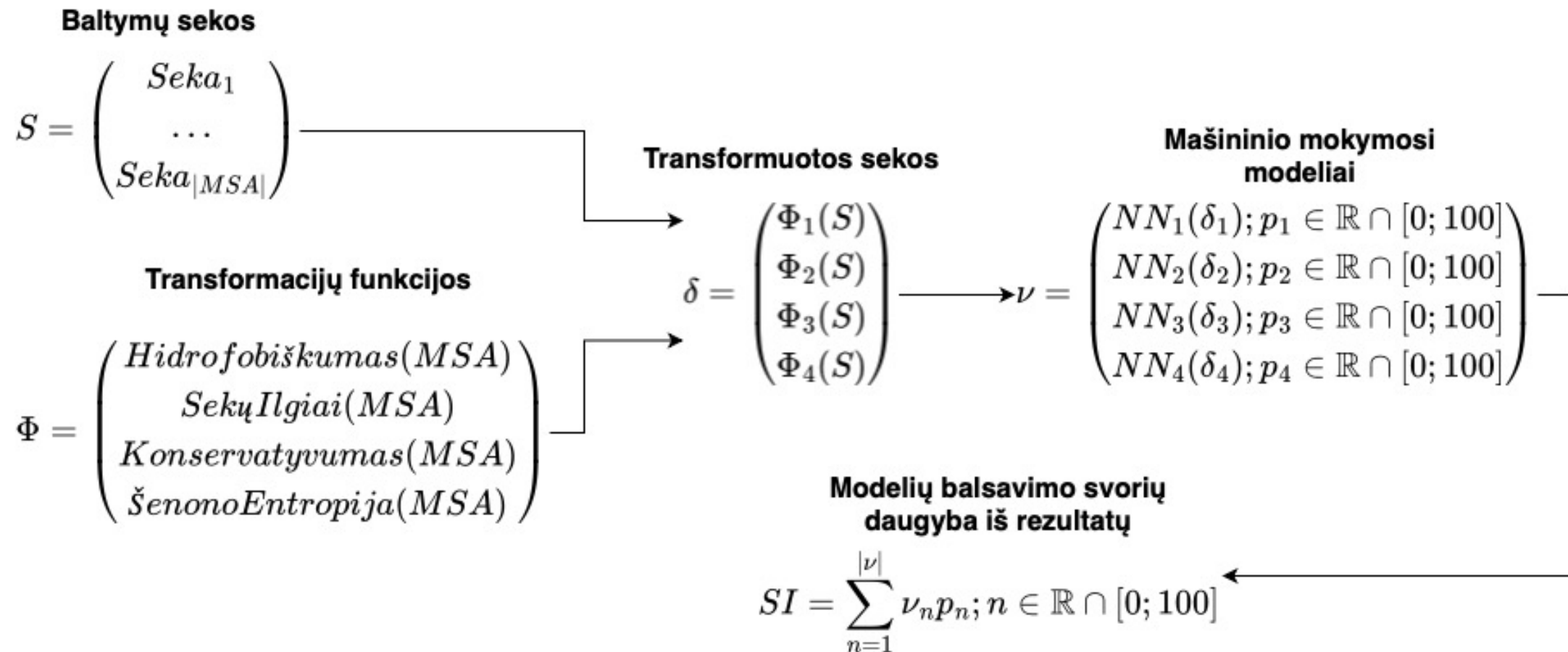
Lentelė 2. 2-as bandymas

Rezultatai (2)



Pav. 5 Rezultatų palyginimas

Pateiktas pasiūlymas



Pav. 6 Pasiūlyta skaičiavimo schema

Rezultatai

- Darbe buvo praplėsta ColabFold programa, leidžiant nurodyti savo pradinių duomenų paieškos serverio adresą. Taip yra leidžiama modifikuoti esamą integraciją su MMSeq2 sprendimu.
- ColabFold praplėstas DIAMOND ir SWIPE programomis.
- Bandymas skaičiuoti suderinamumo indekso reikšmes sekoms buvo nesėkmingas.

Išvados

- Pasiūlyta metrika skaičiuoti suderinamumo indeksą nėra veiksminga.
- Sekų paieškos serveris, kurį naudoja ColabFold, yra per daug apribotas.
- ColabFold per daug riboja pradinio duomenų rinkinio sudarymo nustatymus.
- MMSeq2 nėra universalus sekų paieškos algoritmas.