Baltymų struktūrų modeliavimas trimatėje erdvėje

Paulius Milmantas
Darbo vadovas: doc. dr. Linas Petkevičius

Vilniaus Universitetas
Matematikos ir informatikos fakultetas

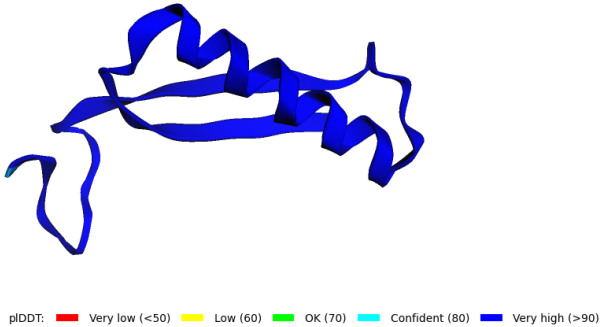
MTD darbo III dalies gynimas

Tiriama sritis (1)

Kas yra tiriama?
Darbe yra tiriami galimi baltymų struktūrų modeliavimo algoritmo AlphaFold2-Multimer 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 [1].

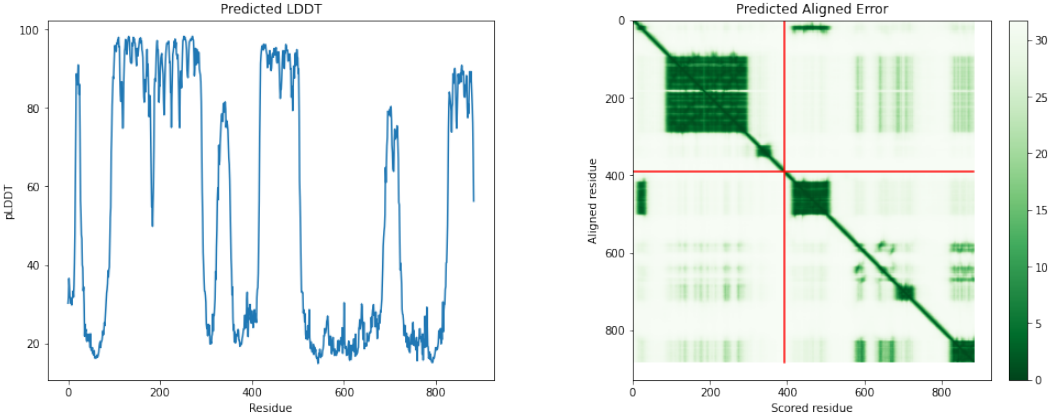
Tiriama sritis (2)

Baltymų seka:
PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASK



pav. 1: Susilankstęs baltymas

Tiriama sritis (3)



pav. 2: Tikslumo metrikos

Hipotezė

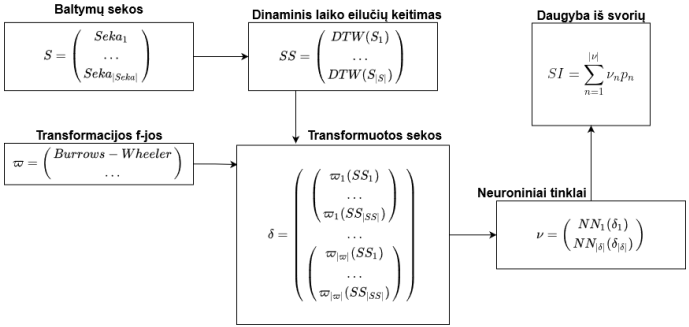
Hipotezė - AlphaFold-Multimer algoritmą galima pagerinti - padidinti išvesties tikslumą, pakeitus MSA sudarymo algoritmą, atsižvelgiant į įvairias baltymų sekų savybes.
MSA - daugybinis sekų palyginys (angl. "Multiple sequence alignment").
Tikslas - Patikrinti hipotezę dėl AlphaFold išvesties tikslumo gerinimo.
Uždavinys - Realizuoti pasiūlytą algoritmą - pagerintą MSA sudarymo algoritmą.

```
CLUSTAL multiple sequence alignment by Kalign (3.3.1)

ENA|BAA20512|BAA20512.1     ATGAGTCTCTCTGATAAGGACAAGGCTGCTGTGAAAGCC-CTATGGGCTAAGATCAGCC-
ENA|CAA23748|CAA23748.1     ATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCC-TGGGGTAAGGTCGGCG-
ENA|CAA24095|CAA24095.1     ATGGTGCTCTCTGGGGAAGACAAAAGCAACATCAAGGCTGCC-TGGGGGAAGATTGGTGG
ENA|CAA28435|CAA28435.1     ATGTCTCTGACCAGGACTGAGAGGACCATCATCCTGTCC-CTGTGGAGCAAGATCTCCA-
```

pav. 3: Sugeneruotas MSA pavyzdys

AlphaFold2 tobulinimas



pav. 4: Tobulinimo pasiūlymas [1]

Prielaida ir pastebėjimai

1. Darbe yra daroma prielaida, kad pagerinus MSA sudarymo algoritmą ir AlphaFold2 tikslumą su mažos apimties MSA struktūra, panašūs rezultatai bus gauti ir su nesutrumpintomis MSA struktūromis.
2. Tyrimo bandymuose pastebėta - egzistuoja tam tikros sekos MSA struktūrose, be kurių nustoja veikti AlphaFold2.

Sekų kritiškumas (1)

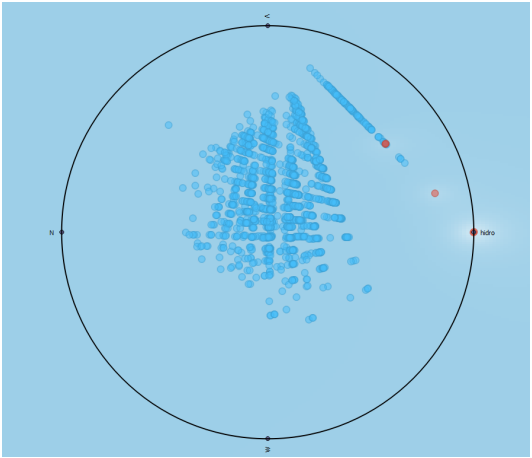
lentelė 1: Spirmano koreliacija

Palyginys	Palyginys	Reikšmė	Paaiškinimas
Kritiškumas	G	0,061	Statistiškai nereikšminga
Kritiškumas	H	0,058	Statistiškai nereikšminga
Kritiškumas	S	0,058	Statistiškai nereikšminga

lentelė 2: Pirsono koreliacija

Palyginys	Palyginys	Reikšmė	Paaiškinimas
Kritiškumas	Hidrofobiškumas	0,128	Silpnas statistinis ryšys
Kritiškumas	G	0,118	Silpnas statistinis ryšys
Kritiškumas	S	0,104	Silpnas statistinis ryšys

Sekų kritiškumas (2)



pav. 5: Klasterių nebuvimas

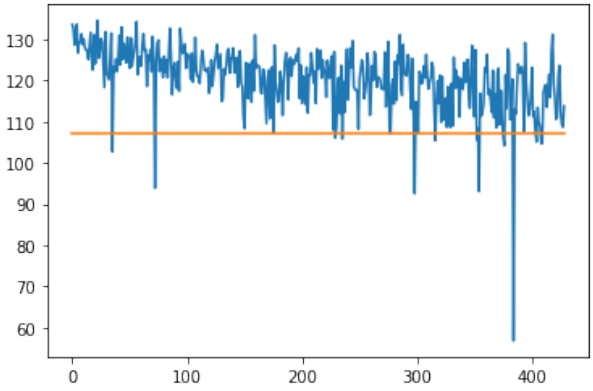
Sekų kritiškumas (3)

lentelė 3: Mašininio mokymosi modeliai

Modelis	AUC	CA	F1	Tikslumas	Grąžinimas (recall)
Neuroninis tinklas	1.00	0.999	0.999	0.999	0.999
Atsitiktinis miškas	1.00	1.00	1.00	1.00	1.00
Medžiai	0.5	0.998	0.997	0.996	0.998
Naive Bayes	1.00	0.821	0.899	0.998	0.821
AdaBoost	1.00	1.00	1.00	1.00	1.00

Sekų atrinkimas

Skaičiavimams pasitelkiant Google Colab platformą ir vykdant skaičiavimą su 429 sekomis, skaičiavimas užtruko 5 minutes.



pav. 6: Atstumai pagal dinaminio programavimo algoritmą

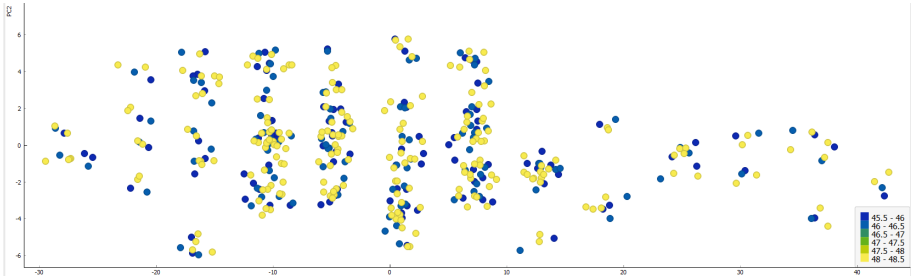
Hidrofobiškumo funkcija

lentelė 4: Mašininio mokymosi modeliai tiriant hidrofobiškumą

Modelis	MSE	RMSE	MAE	R2
Neuroninis tinklas	81,251	9,014	4,673	-38,879
Atsitiktinis miškas	3,340	1,828	1,747	-0,640
AdaBoost	5.596	2.365	2.090	-1.746
SGD	3626352448.36	60219.20	18033.18	-1779874863.34
Tiesinė regresija	4.889	2.211	1.9667	-1.399
Jungtinis (SGD, TR**)	1.735	1.317	1.246	0.148
Jungtinis (GB, AT**)	1.472	1.213	1.113	0.277

** SGD - Stochastinis gradiento nuolydis, TR - tiesinė regresija, GB - gradientų sustiprinimas, AT - atsitiktinis miškas

Komponenčių analizė



pav. 7: Pagrindinių komponentų analizė

Dimensijų mažinimas

pav. 8: Daugiamatis mastelio keitimas, taikant Euklido atstumo skaičiavimą

Rezultatai

- Darbe buvo sukurtas sekų kritiškumo modelis. Šis modelis prognozuoja, ar išmetus tam tikrą seką iš MSA struktūros AlphaFold toliau veiks.
- Darbe buvo sukurtas sekų atrinkimo algoritmas.
- Sukurti mašininio mokymosi modeliai, kurie su < 2 MSE (vidutinė kvadratinė paklaida), geba prognozuoti AlphaFold rezultatų tikslumą.

Išvados

1. Hidrofobiškumo skaičiavimo funkcija grįsti mašininio mokymosi modeliai, su nedideliu duomenų kiekiu, geba prognozuoti sekų išsidėstymo tvarką MSA struktūroje, padidinant AlphaFold tikslumą
2. Sukuriant mašininio mokymosi modelį, buvo įrodyta, kad įmanoma atskirti kritiškas MSA sekas.

Ačiū už dėmesį

Šaltiniai

[1] AlphaFold: a solution to a 50-year-old grand challenge in biology, 2020, https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology