Predikcija neuređenosti proteina korišćenjem velikih jezičkih modela: DisPredict3.0

Anđela Damnjanović

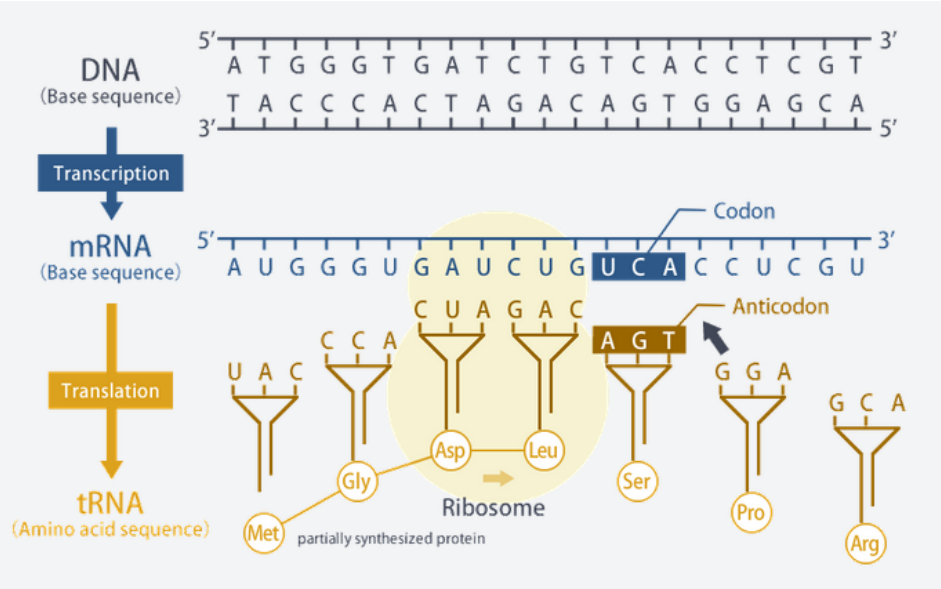
Univerzitet u Beogradu, Matematički fakultet

2. jul 2024.

Sadržaj

Biološki okviri

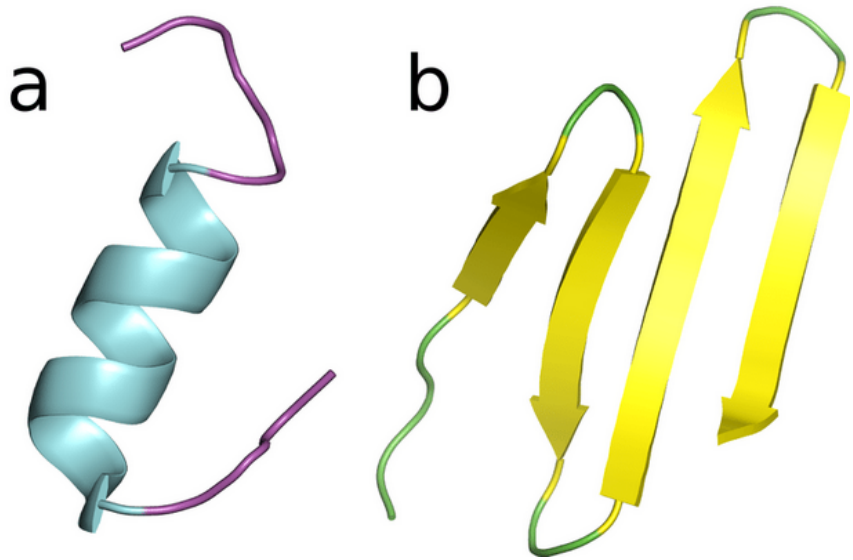
- Proteini
  - definicija
  - uloge



Slika: Put od DNK do proteina [11]

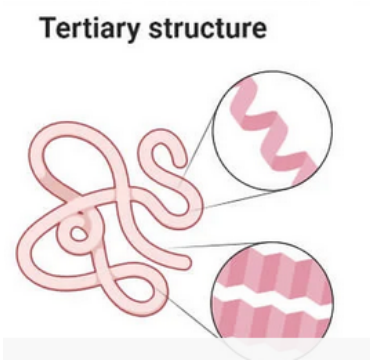
Biološki okviri

- Proteini
  - primarna, sekundarna i tercijarna struktura



Slika: Primeri α-heliksa i β-ravni [8]

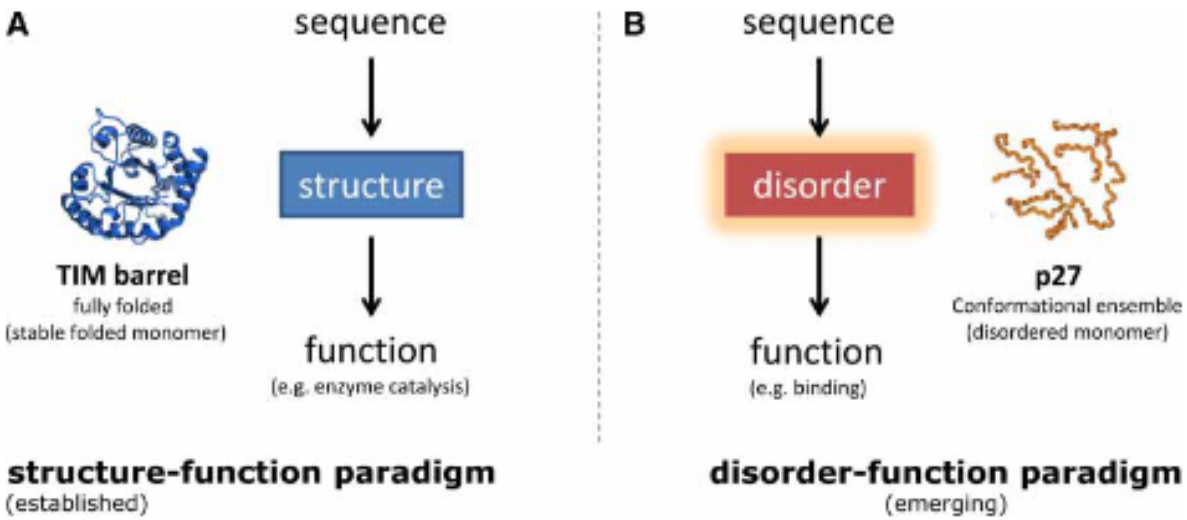
Biološki okviri



Slika: Primer tercijarne strukture proteina [1]

Biološki okviri

- Proteini
  - neuređenost

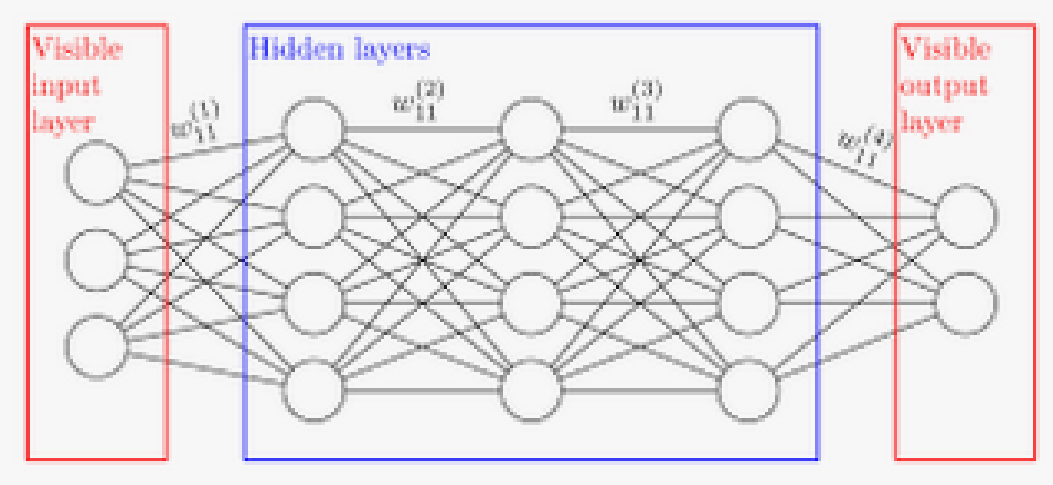


Slika: Primeri uređenog i neuređenog proteina [2]

Veliki jezički modeli

- definicija
- učenje na ogromnim skupovima podataka
- implementacija (neuronske mreže)
- transformeri
- primene
- Evolutionary Scale Modeling (ESM)
  - konvolutivna mreža sa 34 sloja
  - istreniran na preko 86 milijardi AK i 250 miliona proteinskih sekvenci

Veliki jezički modeli



Slika: Primer neuronske mreže [5]

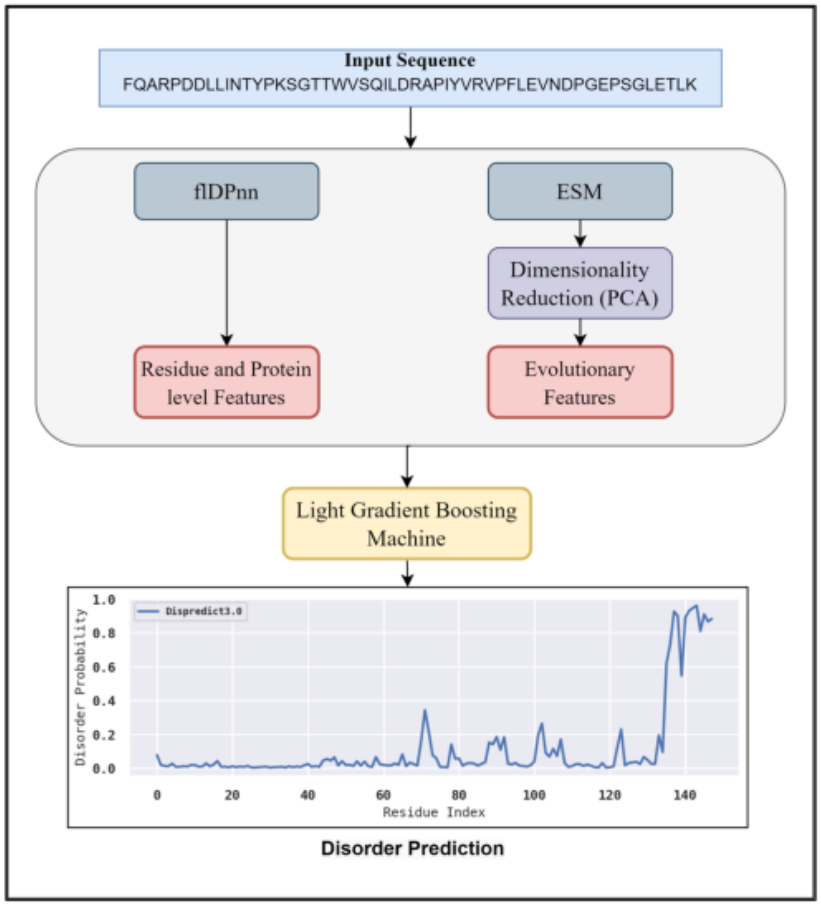
Opis problema

- zadatak: napraviti prediktor neuređenih regiona
- problemi dosadašnjih pristupa
  - tačnost
  - efikasnost
- poznati prediktori
  - AUCpreD
  - SPOT-disorder2
  - ESpritz
  - **flDPnn**
- rešenje problema = DisPredict3.0?

Implementacija

- DisPredict3.0 = fIDPnn + ESM + Light Gradient Boosting Machine (LightGBM)
- fIDPnn – informacije o proteinu na nivou AK
- ESM – predviđa strukturu proteina
  - problem dimenzionalnosti
  - problem ograničene veličine ulaza
- LightGBM – smanjena količina memorije i povećana efikasnost

Implementacija



Slika: Arhitektura DisPredict3.0 [10]

Implementacija

- treniranje modela – DisProt baza
- trening, test i validacija
- nebalansiranost klasa
  - važan izbor pragova i hiperparametara
- prag = 0.382
- broj stabala = 1000

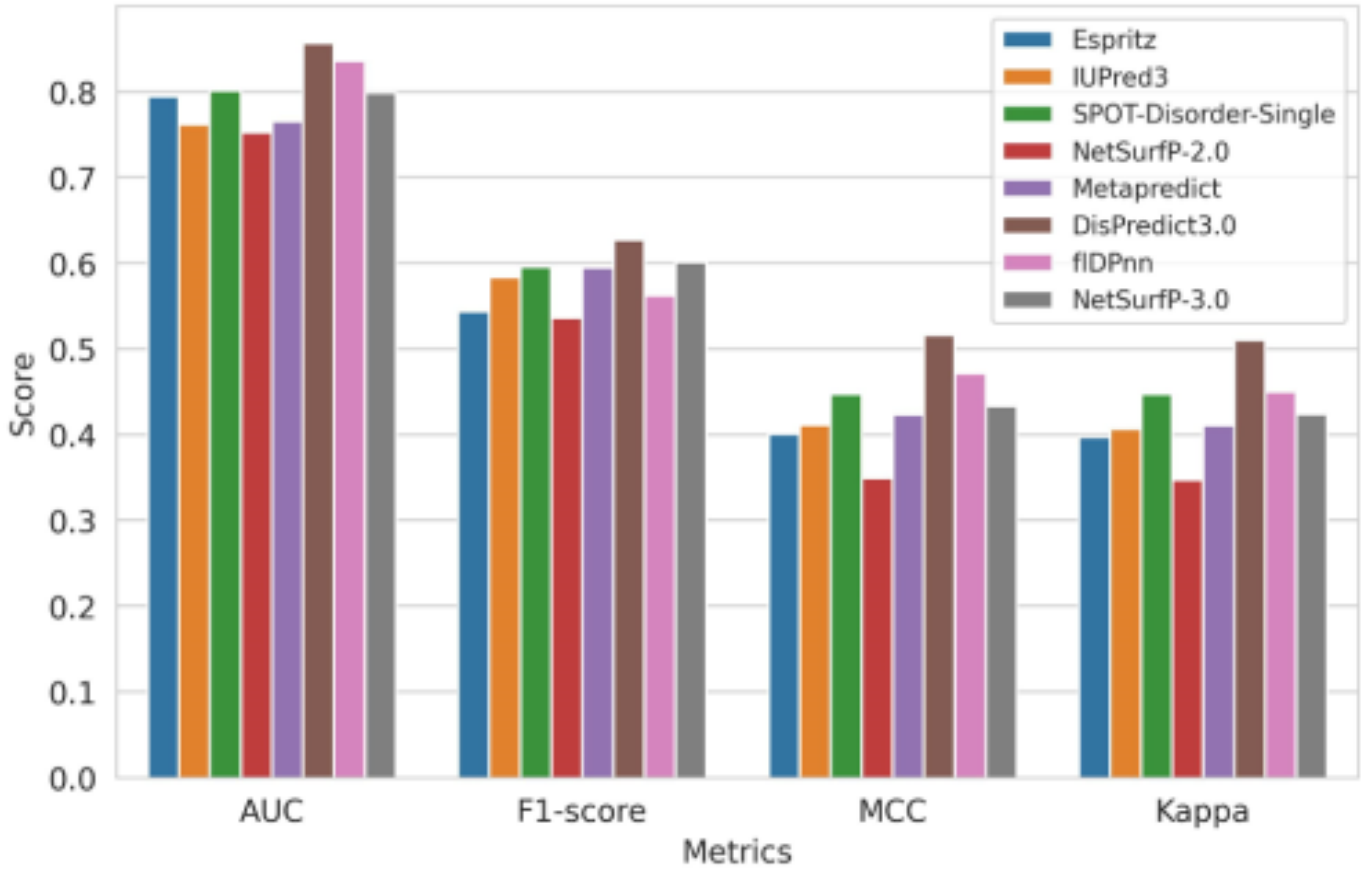
Rezultati

- Mere kvaliteta:
  - AUC
  - f1-skor
  - MCC
  - kapa koeficijent
- najbolji rezultati u svim kategorijama

Metric	Definition
TP	True Positive: Correctly predicted positive samples
TN	True Negative: Correctly predicted negative samples
FP	False Positive: Incorrectly predicted positive samples
FN	False Negative: Incorrectly predicted negative samples
F1-score	$\frac{2TP}{2TP + FP + FN}$
MCC	$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}}$
Kappa	$\frac{2 \times (TP \times TN - FP \times FN)}{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}$

Slika: Računanje metrika [10]

Rezultati



Slika: Poređenje rezultata različitih prediktora [10]

Rezultati

- Ostali rezultati:
  - 1. mesto u predviđanju na No X-ray skupu podataka
  - 6. mesto na PDB skupu podataka
  - bolji od fIDPnn za većinski neuređene sekvence
  - podjednako dobar kao i fIDPnn za skroz neuređene sekvence
  - brže izvršavanje
  - dodatno ubrzanje paralelizacijom

Zaključak

- inovativan pristup
- dobri rezultati
- slaba tačka: dosta vremena se troši na učitavanje modela
- prostor za napredak

Literatura

[1] Sagar Aryal. Protein Structure- Primary, Secondary, Tertiary, and Quaternary, 2022
[2] Madan M. Babu. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease, 2016.
[3] Engelbert Buxbaum. Fundamentals of Protein Structure and Function. Springer, 2015.
[4] Cloudflare. What is a large language model (LLM)?
[5] DeepAI. Hidden Layer

Literatura

[6] elastic. What is a large language model (llm)?
[7] IBM. What are LLMs?
[8] Greg Lever. Large Scale Quantum Mechanical Enzymology
[9] Jeffrey Skolnick and Jacquelyn S. Fetrow. From genes to protein structure and function: novel applications of computational approaches in the genomic era, 2019
[10] Wasi Ul Kabir and Tamjidul Hoque. Dispredict3.0: Prediction of intrinsically disordered regions/proteins using protein language model.
[11] What is protein? Protein Crystal Growth

Github

- preuzeti kod sa: zvaničnog Github naloga
- dodati biblioteku torch unutar dela za zavisnosti u pyproject.toml fajlu
- pokrenuti skript ./install_dependencies.sh
- pokrenuti komande

```
cd script
../.venv/bin/poetry run python Dispredict3.0.py -f "../example/
    sample.fasta" -o "../output/"
```

Docker

- instalirati Docker pomoću sudo apt install docker
- pokrenuti komande

```
sudo groupadd docker
sudo usermod -aG docker ${USER}
sudo chmod 666 /var/run/docker.sock
sudo systemctl restart docker
docker run -ti --name dispredict3.0  wasicse/dispredict3.0:latest
```

- pokrenuti program pomoću komandi


```
sudo groupadd docker
sudo usermod -aG docker ${USER}
sudo chmod 666 /var/run/docker.sock
sudo systemctl restart docker
docker run -ti --name dispredict3.0  wasicse/dispredict3.0:latest
```

Singularity

- instalirati programski jezik Go
- instalirati alat Singularity
- skinuti DisPredict3.0
- pokrenuti program

Singularity

```
sudo yum update -y && \
    sudo yum groupinstall -y 'Development Tools' && \
    sudo yum install -y \
    openssl-devel \
    libuuid-devel \
    libseccomp-devel \
    wget \
    squashfs-tools
// instalacija programskog jezika Go
sudo apt-get update
wget https://go.dev/dl/go1.21.0.linux-amd64.tar.gz
sudo tar -xvf go1.21.0.linux-amd64.tar.gz
sudo mv go /usr/local
export GOROOT=/usr/local/go
export GOPATH=$HOME/go
export PATH=$GOPATH/bin:$GOROOT/bin:$PATH
source ~/.profile

// instalacija alata Singularity
echo 'export GOPATH=${HOME}/go' >> ~/.bashrc && \
    echo 'export PATH=/usr/local/go/bin:${PATH}:${GOPATH}/bin' >>
    ~/.bashrc && \
    source ~/.bashrc
go get -u github.com/golang/dep/cmd/dep
export VERSION=v3.0.3 # or another tag or branch if you like && \
    cd $GOPATH/src/github.com/sylabs/singularity && \
    git fetch

./mconfig && \
    make -C ./builddir && \
    sudo make -C ./builddir install

// pokretanje alata DisPredict3.0
singularity pull dispredict3.sif docker://wasicse/dispredict3.0
singularity run --writable-tmpfs dispredict3.sif

export PATH="/opt/poetry/bin:${PATH}"
source /opt/Dispredict3.0/.venv/bin/activate
python /opt/Dispredict3.0/script/Dispredict3.0.py -f "/opt/
    Dispredict3.0/example/sample.fasta" -o "/opt/Dispredict3.0/
    output/"
```