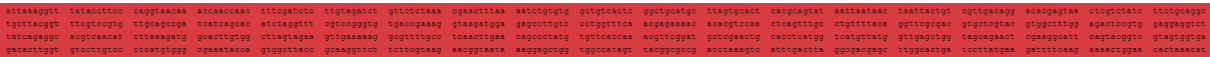
Analiza rada IDPpred

Vladimir Mijić

29. maj 2024.

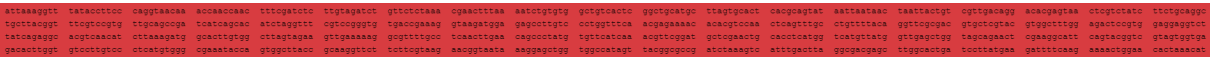


Uvod

Cilj prezentacije: Analiza i razumijevanje metoda i rezultata rada IDPpred

Informacije o radu

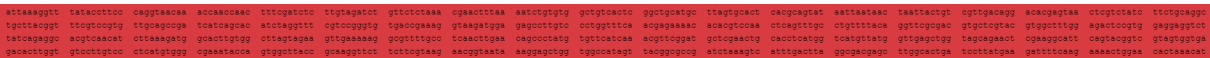
Rad: IDPpred: a new sequence-based predictor for identification of intrinsically disordered protein with enhanced accuracy
Autori: Chaurasiya, D., et al. (2023)
Časopis: *Journal of Biomolecular Structure and Dynamics*
DOI: 10.1080/07391102.2023.2290615



Uvod - Formulacija problema

Intrinsically disordered proteins (IDP)

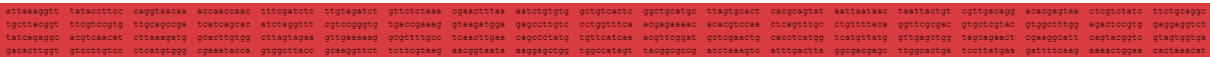
- IDP - Unutrašnje poremećeni proteini, preciznije, unutarmolekulski poremećani proteini
- IDP su proteini koji nemaju fiksnu ili uređenu 3D strukturu



Uvod - Formulacija problema

Intrinsically disordered proteins (IDP)

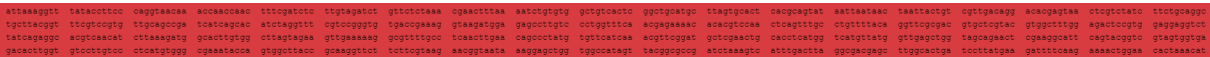
- IDP - Unutrašnje poremećeni proteini, preciznije, unutarmolekulski poremećani proteini
- IDP su proteini koji nemaju fiksnu ili uređenu 3D strukturu
- Sastavljeni od aminokiselina čije veće koncentracije izazivaju poremećaj u strukturi



Uvod - Formulacija problema

Intrinsically disordered proteins (IDP)

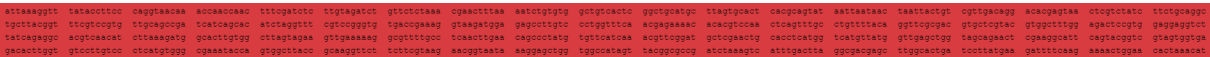
- IDP - Unutrašnje poremećeni proteini, preciznije, unutarmolekulski poremećani proteini
- IDP su proteini koji nemaju fiksnu ili uređenu 3D strukturu
- Sastavljeni od aminokiselina čije veće koncentracije izazivaju poremećaj u strukturi
- Lys, Arg, Glu i Pro su aminokiseline koje se često pojavljuju u IDP proteinima



Uvod - Formulacija problema

Intrinsically disordered proteins (IDP)

- IDP - Unutrašnje poremećeni proteini, preciznije, unutarmolekulski poremećani proteini
- IDP su proteini koji nemaju fiksnu ili uređenu 3D strukturu
- Sastavljeni od aminokiselina čije veće koncentracije izazivaju poremećaj u strukturi
- Lys, Arg, Glu i Pro su aminokiseline koje se često pojavljuju u IDP proteinima
- IDP proteini su ključni u mnogim biološkim procesima



Uvod - Formulacija problema

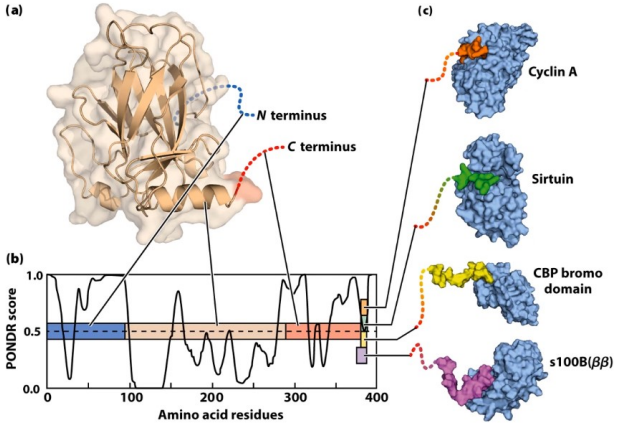
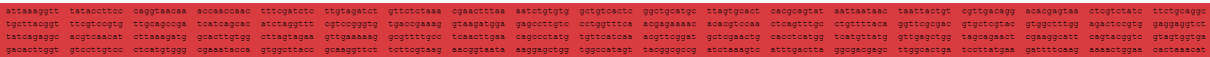


Figure 4-22
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

Uvod - Formulacija problema

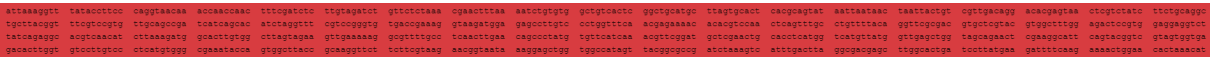
Pitanje koje se istražuje: Može li se razviti efikasan računarski model za predikciju IDP proteina samo na osnovu njihove sekvence aminokiselina?



Uvod - Formulacija problema

Pitanje koje se istražuje:
Može li se razviti efikasan računarski model za predikciju IDP proteina samo na osnovu njihove sekvence aminokiselina?
Ciljevi istraživanja:

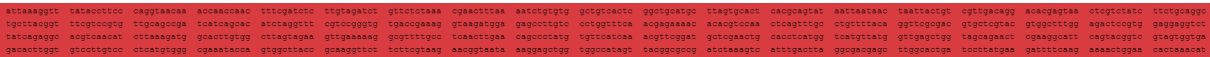
- Razvoj novog prediktora (IDPpred) za identifikaciju IDP proteina
- Poboljšanje tačnosti predikcije u odnosu na postojeće metode
- Fokus na kratkim nestrukturiranim regionima i proteinima koji u potpunosti nemaju uređenu strukturu



IDPpred Metodologija

Koraci u razvoju IDPpred prediktora:

1. **Prikupljanje podataka:**
   - Svi podaci su prikupljeni iz CAID2018 takmičenja
2. **Ekstrakcija karakteristika (osobina) zasnovana na sekvenci:**
   - Predstavljanje sekvenci proteina pomoću vektora
   - Korišćenje različitih metoda za ekstrakciju karakteristika
3. **Dizajn IDPpred modela:**
   - Korišćenje tri klasifikatora za identifikaciju potpuno nestrukturiranih (poremećenih) proteina
   - Uz pomoć back-propagation neuronske mreže *(BPN)*
4. **Evaluacija modela:**
   - Procjena performansi modela na nezavisnom (test) skupu podataka
   - Korišćenjem različitih metrika evaluacije



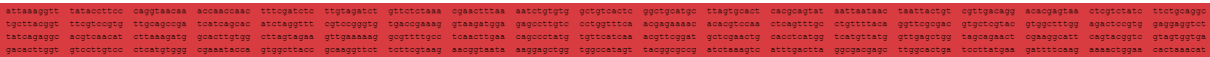
IDPpred Metodologija - Prikupljanje podataka

DisProt baza podataka:

- Ručno kreirana baza podataka IDP proteina
- Sadrži eksperimentalno potvrđene IDP sekvence
- Korišćena verzija *DisProt 2022.12.*
- 142 potpuno nestrukturirana proteina (>95% nestrukturiranih ostataka)
- Licenca: *Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License*

MobiDB baza podataka:

- Baza podataka IDP proteina
- Korišćena za prikupljanje dodatnih podataka o nestrukturiranim proteinima za kontrolni skup
- Licenca: *Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License*



IDPpred Metodologija - Prikupljanje podataka (nastavak)

Kontrolni skup:

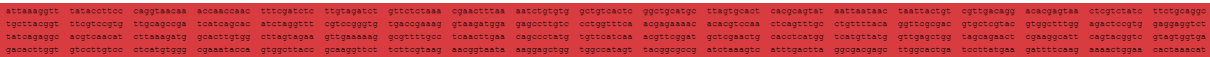
- Korišćeni proteini različitih organizama
- Nestruktuiranost potvrđena metodama:
  - PONDR pool (PONDR-FIT, PONDR-VLXT, PONDR-VSL2)
  - ESpritz (NMR, X-ray, Disprot)

Priprema podataka:

- Korišteni proteini koji imaju više od 30% nestrukturiranih ostataka
- Konverzija u odgovarajući vektor karakteristika *eng. feature vector* (ProtPCV2, ProtIDR, CIDER)
- Nasumična podjela podataka na trening i test skup (70:30)

Skup za testiranje:

- Korišćen CAID skup podataka - izvor DisProt
  - 45 potpuno nestrukturiranih proteina
- ostatak su kontrolni proteini



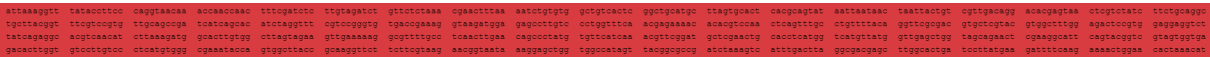
IDPpred Ekstrakcija karakteristika

ProtIDR

- Zasnovana na klasifikaciji aminokiselina kao:
  - **OPaa** (Order Promoting Amino Acids) - aminokiseline koje promovišu uređenost
  - **DPaa** (Disorder Promoting Amino Acids) - aminokiseline koje promovišu neuređenost
- 10-dimenzioni vektor

ProtCV2

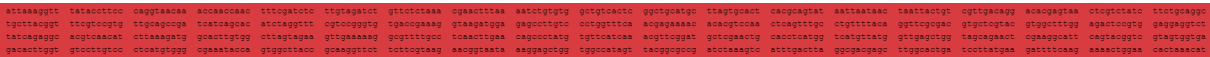
- Koristi 5 fizičko-hemijskih osobina aminokiselina:
  - Hidrofobnost *(eng. Hydrophobe - HC)*
  - Polarnost *(eng. Polarity - PO)*
  - Potencijal naelektrisanja *(eng. Charge Characteristic - CC)*
  - Hidrofobni indeks *(eng. Hydrophobe Index - HI)*
  - Hidropatski indeks *(eng. Hydropathy Index - HY)*
- 50-dimenzioni vektor (10 po osobini)



IDPpred Ekstrakcija karakteristika (nastavak)

CIDER

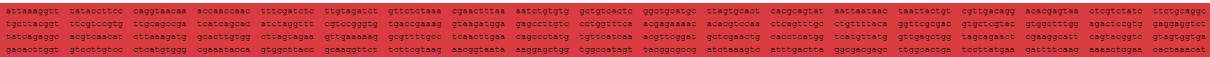
- Postojeći skup od 10 karakteristika dobijenih sa CIDER web servera.
- Uključuje:
  - Odnos negativnog i pozitivnog naelektrisanja (f- i f+)
  - Vrijednosti Das-Pappu skale
  - FCR - fraction of charged residues
  - NCPR - net charge per residue
  - Parametre Kappa, Omega, Sigma, Delta, Max Delta i Hydropathy



IDPpred Model

Ključne osobine:

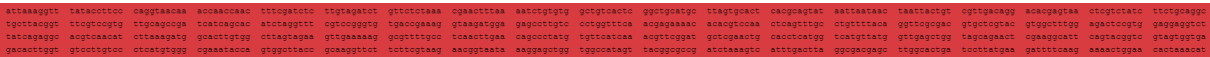
- Koristi tri različita klasifikatora zasnovana na neuronskim mrežama
- Svaki klasifikator koristi drugačiji skup karakteristika izvedenih iz sekvence proteina:
  - ProtPCV2 (50 dimenzija)
  - ProtIDR (10 dimenzija)
  - CIDER (10 dimenzija)
- Finalna predikcija se donosi glasanjem (voting) između tri klasifikatora



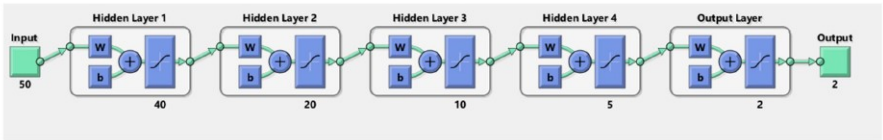
IDPpred Model (nastavak)

Princip glasanja:

- Za oznaku klase 'F disord', izlaz je 1.
- Za oznaku klase 'Control', izlaz je 0.
- Ako je zbir izlaza veći od 2, protein se klasifikuje kao 'F disord', inače 'Control'.



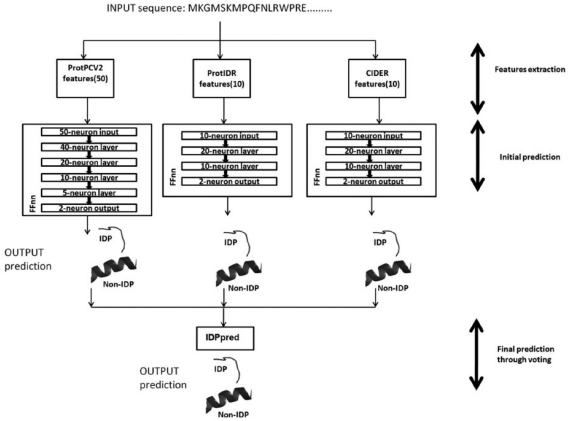
IDPpred Model - BPN arhitektura



Slika: BPN arhitektura: a) 50dim ProtPCV i b) 10dim ProtIDR ili CIDER vektori

atttaaaggtt tatacctttcc caggtaacaa accaaccaac tttcgatctc ttgtagatct gttctctaaa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact cacgcagtat aattaataac taatttactgt cgttgacagg acacgagtaa ctcgtctatc ttctgcaggc
tgcttacggt ttcgtccgtg ttgcagccga tcatcagcac atctaggttt cgtccgggtg tgaccgaaag gtaagatgga gagccttgtc cctggtttca acgagaaaac acacgtccaa ctcagtttgc ctgttttaca ggttcgcgac gtgctcgtac gtggctttgg agactccgtg gaggaggtct
tatcagaggc acgtcaacat cttaaagatg gcacttgtgg cttagtagaa gttgaaaaag gcgttttgcc tcaacttgaa cagcctatg tgttcatcaa acgttcggat gctcgaactg cacctcatgg tcatgttatg gttgagctgg tagcagaact cgaaggcatt cagtacggtc gtagtggtga
gacacttggt gtccttgtcc ctcatgtggg cgaaatacca gtggcttacc gcaaggttct tcttcgtaag aacggtaata aaggagctgg tggccatagt tacggcgccg atctaaagtc atttgacgta ggcgacgagc ttggcactga tccttatgaa gatttttcaag aaaactggaa cactaaacat

IDPpred Model - Workflow



Slika: Dijagram toka rada IDPpred prediktora

atttaaaggtt tatacctttcc caggtaacaa accaaccaac tttcgatctc ttgtagatct gttctctaaa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact cacgcagtat aattaataac taattactgt cgttgacagg acacgagtaa ctcgtctatc ttctgcaggc
tgcttacggt ttcgtccgtg ttgcagccga tcatcagcac atctaggttt cgtccgggtg tgaccgaaag gtaagatgga gagccttgtc cctggtttca acgagaaaac acacgtccaa ctcagtttgc ctgttttaca ggttcgcgac gtgctcgtac gtggctttgg agactccgtg gaggaggtct
atcagaggcc acgtcaacat cttaaagatg gcacttgtgg cttagtagaa gttgaaaaag gcgttttgcc tcaacttgaa cagcctatg tgttcatcaa acgttcggat gctcgaactg cacctcatgg tcatgttatg gttgagctgg tagcagaact cgaaggcatt cagtacggtc gtagtggtga
gacacttggt gtccttgtcc ctcatgtggg cgaaatacca gtggcttacc gcaaggttct tcttcgtaag aacggtaata aaggagctgg tggccatagt tacggccgcg atctaaagtc atttgactta ggcgacgagc ttggcactga tccttatgaa gatttttcaag aaaactggaa cactaaacat

IDPpred Evaluacija

Za procjenu prediktivnih performansi modela IDPpred i drugih prediktora korišćene su sljedeće metrike:

Osjetljivost (Sensitivity):

$$\text{Sensitivity} = \frac{TP}{N_{\text{dis}}}$$

Tačnost (Accuracy):

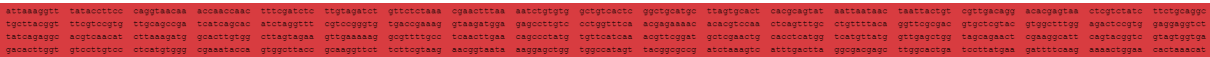
$$\text{Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$

Selektivnost (Selectivity):

$$\text{Selectivity} = \frac{TP}{TP + FP}$$

Specifičnost (Specificity):

$$\text{Specificity} = \frac{TN}{N_{\text{ord}}}$$



IDPpred Evaluacija (nastavak)

Za procjenu prediktivnih performansi modela IDPpred i drugih prediktora korišćene su sljedeće metrike:

F-mjera (F-measure):

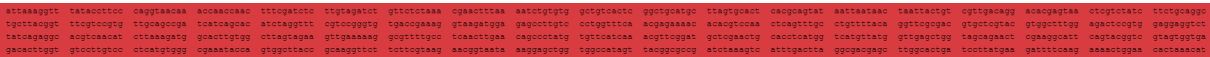
$$\text{F1} = \frac{2 \cdot \text{Sensitivity} \cdot \text{Selectivity}}{\text{Sensitivity} + \text{Selectivity}}$$

Matthews correlation coefficient (MCC)

Ocjena (Sw):

$$S_w = \text{Sensitivity} + \text{Specificity} - 1$$

TP: broj tačno pozitivnih (true positives)
TN: broj tačno negativnih (true negatives)
FP: broj lažno pozitivnih (false positives)
FN: broj lažno negativnih (false negatives)
N_{dis}: ukupan broj nestrukturiranih proteina
N_{ord}: ukupan broj strukturiranih proteina

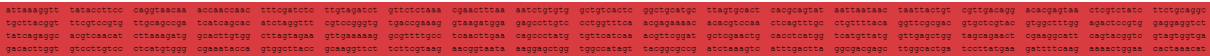


IDPpred Rezultati

	Performance evaluation parameters*									
Predictors	TP	TN	FP	FN	TNR	TPR	BAC	MCC	F1-s	PPV
IDPpred	**37**	**515**	**92**	**8**	**0.8484**	**0.8222**	**0.8353**	**0.4267**	**0.4252**	**0.2868**
CIDER	**35**	**492**	**115**	**10**	**0.7777**	**0.8105**	**0.7935**	**0.3543**	**0.2333**	**0.3590**
flDPnn	26	585	16	19	0.973	0.578	0.776	0.569	0.598	0.619
ProtIDR	**32**	**488**	**119**	**13**	**0.71**	**0.80**	**0.755**	**0.3295**	**0.3316**	**0.2119**
ProtPCV2	**33**	**431**	**176**	**12**	**0.7333**	**0.710**	**0.7216**	**0.2408**	**0.2598**	**0.157**

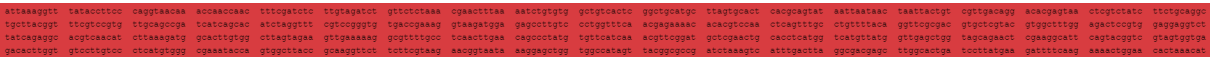
*TN, true negatives count; TP, true positives count; FN, false negatives count; FP, false positives count; F1-s, F1-score; TNR, true negative rate, specificity; TPR, true positive rate, recall; PPV, positive predictive value, precision; BAC= (sensitivity + specificity)/2, balanced accuracy for prediction of fully disordered proteins. Proteins with disorder prediction or disorder annotation covering at least 95% of the sequence are considered fully disordered. Predictors are sorted by their BAC.

Slika: Rezultati evaluacije IDPpred prediktora



Zaključak

- IDPpred je novi i efikasni prediktor za identifikaciju IDP proteina
- Koristi kombinaciju različitih karakteristika i algoritama mašinskog učenja
- Postigao je bolje performanse od postojećih prediktora
- Ima potencijal da unaprijedi istraživanja u oblasti IDP proteina uz kombinaciju sa drugim metodama

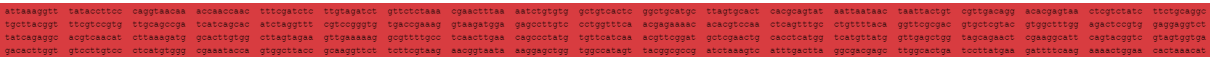


Zaključak

- IDPpred je novi i efikasni prediktor za identifikaciju IDP proteina
- Koristi kombinaciju različitih karakteristika i algoritama mašinskog učenja
- Postigao je bolje performanse od postojećih prediktora
- Ima potencijal da unaprijedi istraživanja u oblasti IDP proteina uz kombinaciju sa drugim metodama

Aktuelnosti

- AlphaFold3 bi mogao biti primjenjen za ovaj problem - Abramson et al. (May 2024)
- Objavljen je flDPnn2 - Wang et al. (May 2024)



Hvala na pažnji!

Reference

- Chaurasiya, D., et al. (2023). IDPpred: A new sequence-based predictor for identification of intrinsically disordered proteins.
- Necci, M., et al. (2021). Critical assessment of protein intrinsic disorder prediction. *Nature Methods*, 18, 472-481.
- Lotthammer, J.M., et al. (2024). Direct prediction of intrinsically disordered protein conformational properties. *Nature Methods*, 21, 465-476.
- Nelson, D.L., & Cox, M.M. (2012). *Lehninger Principles of Biochemistry*.