

# Analiza rada IDPpred

Vladimir Mijić

29. maj 2024.

attaaagggt	tataccttcc	caggtaacaa	accaaccaac	tttcgatctc	ttgtagatct	gttctctaaa	cgaactttaa	aatctgtgtg	gtgtctactc	ggctgcatgc	ttagtgcact	cacgcagtat	aatttaataac	taattactgt	cgttgacagg	acacgagtaa	ctcgtctatc	ttctgacggc
tgcttaacgt	ttcgtccgtg	ttgcagccga	tcctcagcac	atctagggtt	cgtccgggtg	tgaccgaaag	gtaagatgga	gagcctgtgc	cctgggttcc	acgagaaaaac	acacgtccaa	ctcagttttg	ctgtttttaca	ggttcgcgac	gtgctcgtac	gtgcttttgg	agactccgtg	gaggagggtc
tatcagaggc	acgtcaacct	cttaagatg	gcacttgtgg	cttagtagaa	gttgaaaaag	gogttttgcc	tcaacttgaa	cagccctatg	tgttcatcaa	acgttcggat	gotcgaactg	caacctcatg	tcactgttatg	gttgagctgg	tagcagaact	cgaaggcatt	cagtaacggt	gtagtgggtc
gcacttgggt	gtccttgtcc	ctcatgtggg	cgaataacca	gtggcttaac	gcaaggttct	tcttcgttaag	aacggttaata	aaggagctgg	tggccatagt	tacggcgccg	atctaaaagtc	atttgactta	ggcgacgagc	ttggcaactga	tccttatgaa	gattttcaag	aaaaactgaa	cactaaacca

# 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

attaaaggtt tataccttcc caggttaacaa accaaccacac tttagatctc ttgtagatct gtctctataa cgaactttta aatctgtgtg gctgtcactc ggctgcctgc ttagtgcact cagcgagtat aattaataac taattactgt cgttgacagg acacgagtaa ctgctctatc ttctgcaggc  
tgcttaaggt ttgctcogtg ttgcagccga tcatcagcac atctaggttt cgtccgggtg tgacgcgaag gtaagatgga gagccttgtc cctggtttca acgagaaaaa acacgtccaa ctacgtttgc ctgttttaca ggttcgcgac gtgctcgtac gtgcctttgg agactcogtg gaggaggtct  
tatcagaggo acgtcaacat cttaaagatg gcactttggy cttagtagaa gttgaaaaag tcaacttgaa cagccctatg tgttcaccaa acgttcggat gctogaactg caacctcatg tcatgttatg gttgagctgy tagcagaact cgaaggcatt cagtaagggtc gtatgggtga  
gacacttggt gtctttgtcc ctcatgtggy cgaataacca gtggttacc gcaaggttct tcttctgaag aacggttaata aaggagctgy tggccatagt taaggcgccg atctaaagtc atttgaatta ggccagcaga ttggcaactga tctttatgaa gattttcaag aaaaactgaa caactaaaca

# 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

attaaagggt	tataccttcc	caggtaacaa	accaaccaac	tttogatctc	ttgtagatct	gttctctaaa	cgaactttaa	aatctgtgtg	gtgttcaact	ggctgcatgc	ttagtgcact	cacgcagtat	aatttaataac	taattactgt	cgttgacagg	acacagataa	cttgtctatc	ttctgacagg
tgcttaacgt	tttgtccgtg	ttgcagccga	tcatacgcac	atctaggttt	cgtccgggtg	tgacgcgaag	gtaagatgga	gagccttgct	cctgggttca	acgcagaaaac	acacgtccaa	ctcagtttgc	ctgtttttaca	ggttcgcgac	gtgtcgtgac	gtggctttgg	agactcogtg	gaggaggtct
tatcagaggg	acgtcaacat	cttaaaagatg	gcacttgctg	cttagtagaa	gttgaaaaag	gcgtttttgcc	tcaacttgaa	cagccctatg	tggttcatcaa	acgttcggat	gtctgaactg	caacctcatg	tcattgttatg	gttgagctgg	tagcagaact	cgaaggcatt	cagtacggctc	gtagtgggtga
gacacttggt	gtccttgtcc	ctcatgtggg	cgaataacca	gtggcttaac	gcaaggttct	tctttgttaag	aacggttaata	aaggagctgg	tggccatagt	taaggcgccg	atctaaaagtc	atttgactta	ggcgacgagc	ttggcaactga	tccttatgaa	gattttcaag	aaaactgcaa	cactaaacac

# Uvod - Formulacija problema

## Intrinsically disordered proteins (IDP)

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

attaaagggt	tataccttcc	caggtaacaa	acaaaccaac	tttogatctc	ttgtagatct	gttctctaaa	cgaactttaa	aatctgtgtg	gtgttcaact	ggctgcatgc	ttagtgcact	cacgcagtat	aatttaatac	taattactgt	cgttgacagg	acacagataa	cttgtctatc	ttctgacagg
tgcttaacgt	tttgtccgtg	ttgcagccga	tcatacagac	atctaggttt	cgtccgggtg	tgacccgaag	gtaagatgga	gagccttgct	cctgggttca	acgagaaaa	acacgtccaa	ctcagtttgc	ctgtttttaca	ggttcgcgac	gtgtctgtac	gtggcttttg	agactcogtg	gaggaggtct
tatcagaggg	acgtcaacat	cttaaaagatg	gcacttctgg	cttagtagaa	gttgaaaaag	gcgtttttgoc	tcaacttgaa	cagccctatg	tggttcaatca	acgttcggat	gtctgaactg	cacctcactg	tcactgttatg	gttgagctgg	tagcagaact	cgaaggcaatt	cagtacgggtc	gtagtgggtga
gacacttggt	gtccttctcc	ctcactgtggg	cgaataacca	gtggcttacc	gcaaggtttct	tctttgttaag	aacggttaata	aaggagctgg	tggccatagt	taaggcgccg	atctaaaagtc	atttgactta	ggcgcagcagc	ttggcaactga	tccttatgaa	gattttcaag	aaaactgga	cactaaacac

# Uvod - Formulacija problema

## Intrinsically disordered proteins (IDP)

- IDP - Unutrašnje poremećeni proteini, preciznije, unutar molekularski poremećeni 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

attaaagggt	tataccttcc	caggtaacaa	acaaaccaac	tttgcgtctc	ttgtagatct	gtttctctaa	cgaactttaa	aattctgtgt	gtgtgcactc	ggctgcactc	ttagtgcact	cacgcagtat	aatttaatac	taattactgt	cgttgacagg	acacagagta	ctcgtctatc	ttctgcaggc
tgcttaacgt	ttcgtccgtg	ttgcagccga	tcacagcac	atctaggttt	cgtccgggtg	tgacgaaag	gtaagatgga	gagcctgtgc	cctgggttca	acgagaaaa	acacgtccaa	ctcagtttgc	ctgtttttac	ggttcgcgac	gtgctcgtac	gtggttttgg	agactcogtg	gaggaggtct
tatcagaggg	acgtcaacat	cttaaaagatg	gcacttctgg	cttagtagaa	gttgaaaaag	gcgtttttgc	tcaacttgaa	cagccctatg	tgcttcacaa	acgttcggat	gttcgaaactg	cacctcactg	tcactgttatg	gttgagctgg	tagcagaact	cgaaggcaatt	cagtacgggtc	gtagtgggtg
gacacttggt	gtccttctcc	ctcactgtggg	cgaataacca	gtggtctaac	gcaaggtttct	tcttctgtaag	aacggttaata	aaggagctgg	tggtcctagt	taaggcgcgc	atctaaaagtc	atttgactta	ggcgcagagc	ttggcaactga	tccttatgaa	gattttcaag	aaaactgga	cactaaacac

# 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

attaaggtt	tataccttc	caggttaaca	accaaccaac	tttogatctc	ttgtagatct	gtttctotaaa	cgaactttta	aattctgtgtg	gtgtgcactc	ggctgcactc	ttagtgcact	cacgcagtat	aattaataac	taattactgt	cgttgacagg	acaagagtaa	ctcgtctatc	ttctgaggc
tgcttaacgt	ttgttcogtg	ttgcagcoga	tcatacagcac	atctaggttt	cgtccgggtg	tgacogaaag	gtaagatgga	gagccttgct	cctggtttca	acgagaaaac	acaagtcocaa	ctcagtttgc	ctgtttttaca	ggttcogcac	gtgctcgtac	gtggcttttg	agactcogtg	gaggaggtct
tatcagaggo	acgtcaacat	cttaagatg	gcacttgctg	cttagtagaa	gttgaaaaag	gogttttgoc	tcaacttgaa	cagccttatg	tgttcatcaa	acgttcggat	gtcogaactg	cacctcactg	tcattgttatg	gttgagctgg	tagcagaact	cgaaggcaatt	cagtacggctc	gtagtgggtga
gacacttggt	gtccttgctc	ctcatgtggg	cgaataacca	gtggotttacc	gcaaggttct	totttctgaag	aacggttaata	aaggagctgg	tgcccatagt	taaggcgcgc	atctaaagtc	atttgactta	ggcgcagcgc	ttggcactga	tccttatgaa	gattttcaag	aaaactgcaa	cactaaacac

# Uvod - Formulacija problema

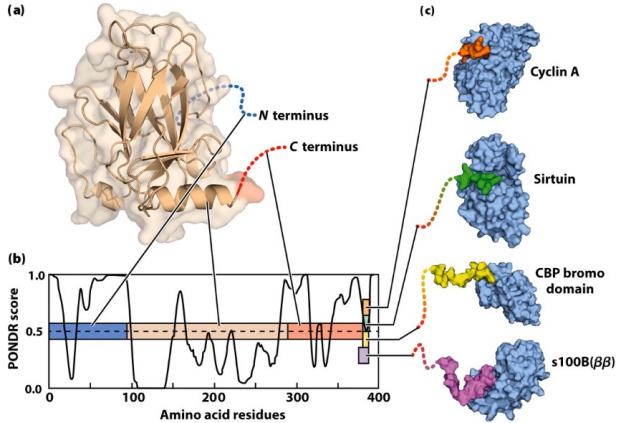


Figure 4-22  
Lehninger Principles of Biochemistry, Seventh Edition  
© 2017 W. H. Freeman and Company

atataagggt tatacatttc caggttaada accaaccaac ttctgatctc ttgtagatct gtctctataa cgaactttta aatctgtgtg gctgtcaact ggcctgcatt ttagtgcaat cagcgagtat aattaataac taattactgt cgttgacagg acacagatga ctgtctatct ttctgcaggc  
tgcttaacgt ttgctcogtg ttgaagocga tcatcagcac atctaggttt cgtccgggtg tgacocgaag gtaagatgga gagcctgtgc cctggtttca ctcggtttca ctcggtttca ctcggtttca ctcggtttca ctcggtttca  
tatcagaggo acgtcaacet cttaaaagtg gaactttgtg cttagtagaa gttagaaag gctgttttgc tcaacttgaa cagcctatg ttgttcaaaa acgttcggat gctgaagactg atctaaagtc atttgactta ggcgcagcgc ttggcaactg tcttatgaa gattttcaag aaaaactgaa cactaaacac  
gacacttggt gtctttgtcc ctcatgtggg cgaataacca gtggcttaac gcaaggttct tcttctgaag aacggttaata aaggaagctgg tggccatagt taaggcgcgc atctaaagtc atttgactta ggcgcagcgc ttggcaactg tcttatgaa gattttcaag aaaaactgaa cactaaacac

# Uvod - Formulacija problema

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

attaaagggt	tataccttcc	caggtaacaa	acaaaccaac	tttcgatctc	ttgtagatct	gttctctaaa	cgaactttaa	aatctgtgtg	gtgttcactc	ggctgcatgc	ttagtgcact	cacgcagtat	aatttaatac	taattactgt	cgttgacagg	acacgagtaa	cttgtctatc	ttctgagggc
tgcttaacgt	tttgtccgtg	ttgaagccga	tcatacgaac	atctagggtt	cgtccgggtg	tgacccgaag	gtaagatgga	gagccttgtc	cctgggttca	acgagaaaaa	acacgtccaa	ctcagtttgc	ctgtttttaca	ggtttccgac	gtgtccgtac	gtggttttgg	agactccgtg	gaggagggtc
tatcagaggo	acgtcaacct	cttaaaagatg	gcacttgtgg	cttagtagaa	gttgaaaaag	gogttttgoc	tcaacttgaa	cagccctatg	tgttcaatca	acgttcggat	gtctaaaagt	atctgaactg	tttgacttta	ggcgcagcgc	ttggcaactga	tccttatgaa	gatttttcaag	aaaaactgaa
gcacttgggt	gtccttgtcc	ctcatgtggg	cgaataacca	gtggcttaac	gcaaggttct	tcttcgttaag	aacggttaata	aaggagctgg	tggccatagt	tacggcgcgc	atctaaaagt	atttgactta	ggcgcagcgc	ttggcaactga	tccttatgaa	gatttttcaag	aaaaactgaa	cactaaacaa



# 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

attaaagggt tataccttcc caggtaacaa accaaccaac ttctgatctc ttgtagatct gtctctctaaa cgaactttta aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact cagcagttat aatttaatac taattactgt cgttgacagg acacagatga ctgtctctatc ttctgacggc  
tgcttaacgt ttgtcccggt ttgaagcaga tcatcagcac atctaggttt cgtccgggtg tgaccgaaag gtaagatgga gagcctgtgc cctgggttca gctgtcactc acacgtccaa ctacgtttgc ctgtttttaca ggttcgcgac gtgtcgttac gtgcttttg gtgactccgtg  
tatcagagga acgtcaacct cttaaaagatg gaacttggg cttagtagaa gttgaaaaag gogttttgoc tcaacttgaa cagccctatg tgttcatcaa acgttcggat gctcgaactg toatgttatg gttgagctgg tagcagaact cgaaggcatt cagtaacggtc gtagtggtga  
gacacttggt gtctttgtcc ctcatgtggg cgaataacca gtggtttacc gcaaggttct totttogtaag aacggttaata aaggaagctgg tggccatagt tacggcgccg atctaaaagtc atttgactta ggcgcagcago ttggcactga tctttatgaa gattttcaag aaaaactggaa cactaaacaa

# 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

attaaagggt tataccttcc caggtaacaa accaaccaac ttctgatctc ttgtagatct gtctctctaaa cgaactttaa aatctgtgtg gctgtcactc ggcctgatgc ttagtgcact cagcagttat aatttaatac taattactgt cgttgacagg acacagatga ctgtctctac ttctgcaggc  
tgcttaacgt ttgtccogtg ttgcagcoga tcatcagcac atctaggttt cgtccgggtg tgaccgaaag gtaagatgga gagccttgtc cctggtttca acacgtccaa ctacgttttg cgtttttaca ggttcogac gtgtcogtac gtggtttgg  
tatcagaggo acgtcaacat cttaaagatg gaactttggy cttagtagaa gttagaaaag gogttttgoc tcaacttgaa cagccttatg tgttcatcaa acgttcoggt gctogaactg caactcctgg tcatgttatg gttgagctgy tagcagaact cgaaggcatt cagtaocggt gtagtggtga  
gacacttggt gtctttgtcc ctcatgtggg cgaataacca gtggtttacc gaaaggttct tctttgtaag aacggttaata aaggagctgy tggccatagt tacggcgccg atctaaaagtc atttgactta ggcgcagago ttggcaactga tctttatgaa gattttcaag aaaactggaa cactaaacaa

# 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*

attaaagggt	tatacctttc	caggtaacaa	accaaccaac	tttcgatctc	ttttagatct	gtttcttaaa	cgaactttaa	aatctgtgtg	gtgttcactc	ggctgcattc	ttagtgcact	cacgcagtat	aatttaatac	taattactgt	cgttgacagg	acacagataa	cttgtctatc	ttctgaggc
tgcttaacgt	tttgtccgtg	ttgacagcga	tcactagcac	atctaggttt	cgtccgggtg	tgaccgaaag	gtaagatgga	gagcctgtgc	cctgggtttc	acagagaaaa	acacgtccaa	ctcagtttgc	ctgtttttac	ggtttcgac	gtgtcgttac	gtggttttgg	agactccgtg	gaggaggtct
tatcagaggo	acgtcaaacct	cttaaaagatg	gcactttgtg	cttagtagaa	gttgaaaaag	gogtttttgc	tcaacttgaa	cagccctatg	tgttcatcaa	acgttcggat	gctcgaacctg	caacctcatg	tcactgttatg	gttgagctgg	tagcagaact	cgaaggcatt	cagtaacggtc	gtagtgggtc
gcactttggt	gtcctttgtc	ctcactgtgg	cgaataacca	gtggcttaac	gcaaggtttc	tctttgttaag	aacggttaata	aaggagctgg	tggccatagt	taaggcgccg	atctaaaagtc	atttgactta	ggcgcagagc	ttggcaactga	tccttatgaa	gattttcaag	aaaactggaa	cactaaacac

# IDPpred Metodologija - Prikupljanje podataka (nastavak)

## Kontrolni skup:

- Korišćeni proteini različitih organizama
- Nestrukturiranost 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

attaaagggt tataccttcc caggtaacaa accaaccaac ttctgatctc ttctagatct gtctctataa cgaactttaa aatctgtgtg gctgtcactc ggctgcctgc ttagtgcact cagcagtat aattataaac taattactgt cgttgacagg acacagatga ctgtctatcc ttctgaggc  
tgcttaacgt ttctgcoctg ttgacagcga tctatagcgc atctaggttt cgtccgggtg tgacgcgaag gtaagatgga gagcctgtgc cctggtttca acagagaaac ctaagtttgc ctgttttaca ggttgcgcac gtgctcgtac gtgcttttg agactcctg gagaggtct  
tatcagaggo acgtcaacat cttaaaagtg gaactgttgg cttagtagaa gttgaaaaag gogttttgoc tcaacttgaa cagcctatg tgcttcaaaa acgttcggat gctcgaactg atctaaaagc atttgactta ggcgcagcgc ttggcactga tcttatgaa gattttcaag aaaaactgaa cactaaaca  
gacacttggt gtccttggcc ctcatgtggg cgaataacca gtggcttacc gcaaggttct tctttgtaag aacggttaata aaggaactgg tggccatagt taccggcgcc atctaaaagc atttgactta ggcgcagcgc ttggcactga tcttatgaa gattttcaag aaaaactgaa cactaaaca

# 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)

attaaagggt	tataccttcc	caggtaacaa	accaaccaac	tttcgatctc	ttttagatct	gttctctaaa	cgaactttaa	aatctgtgtg	gtgtctactc	ggctgcatgc	ttagtgcact	cacgcagtat	aatttaatac	taattactgt	cgttgacagg	acacagataa	ctgtctctac	ttctgcaggc
tgcttaacgt	tttgtccgtg	ttgacagcga	tcctcagcac	atctaggttt	cgtccgggtg	tgaccgaaag	gtaagatgga	gagccctgtc	ccgtggttca	acgagaaaaa	acacgtccaa	ctcagttttg	ctgtttttaca	ggttcgcgac	gtgtcgtgtac	gtggttttgg	agactccgtg	gaggaggtct
tatcagaggg	acgtcaaacct	cttaaaagatg	gcacttgtgg	cttagtagaa	gttgaaaaag	gogtttttgc	tcaacttgaa	cagccctatg	tgttcactca	acgttcggat	gotcgaactg	caacctcatg	tcactgttatg	gttgagctgg	tagcagaact	cgaaggcatt	cagtaacggtc	gtagtgggtc
gcacttgggt	gtccttgtcc	ctcatgtggg	cgaataacca	gtggcttaac	gcaaggttct	tctttgttaag	aacggttaata	aaggagctgg	tggccatagt	tacggcgccg	atctaaaagt	atttgactta	ggcgcagcgc	ttggcactga	tccttatgaa	gattttcaag	aaaactggaa	cactaaacac

# 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

attaaagggt	tataccttcc	caggtaacaa	acaaaccaac	tttcgatctc	ttgtagatct	gtctctctaaa	cgaactttaa	aatctgtgtg	gtgtctactc	ggctgcatgc	ttagtgcact	cacgcagtat	aatttaatac	taattactgt	cgttgacagg	acacgagtaa	ctgtctctatc	ttctgagggc
tgcttaacgt	tttgtccgtg	ttgacagcga	tcactagcac	atctaggttt	cgtccgggtg	tgacccgaag	gtaagatgga	gagccttgtc	cctgggtttca	acgagaaaaac	acacgtccaa	ctcagtttgc	ctgtttttaca	ggtttccgac	gtgtctgtac	gtggttttgg	agactccgtg	gaggagggtc
tatcagaggo	acgtcaacct	cttaaaagtg	gcacttgtgg	cttagtagaa	gttgaaaaag	gcgtttttgac	tcaacttgaa	cagccctatg	tgttcaatcaa	acgttcggat	gctcgaaactg	caacctcatg	tcactgttatg	gttgagctgg	tagcagaact	cgaaggcaatt	cagtaacggtc	gtagtgggtg
gcacttgggt	gtccttgtcc	ctcatgtggg	cgaataacca	gtggcttaac	gcgaagttct	tctttgttaag	aaaggttaata	aaggaagctgg	tggccatagt	tacggcgccg	atctaaaagtc	atttgactta	ggcgacagagc	ttggcaactga	tccttatgaa	gatttttcaag	aaaaactggaa	cactaaacac

# 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

attaaagggt tataccttcc caggtaacaa accaaccaac ttctgatctc ttgtagatct gtctctataa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact cagcagttat aatttaatac taattactgt cgttgacagg acacagatga ctgtctatcc ttctgacggc  
tgtttaacgt ttgtccogtg ttgaagocga tcatcagcac atctaggttt cgtccgggtg tgaccgaaag gtaagatgga gagcctgtgc cctgggttca acagtaaaaa acactgttgc ctactgttgc agactccgtg gaggaggtct  
tatcagaggo acgtcaacct cttaaaagtg gaacttgttg cttagtagaa gttgaaaaaa gogttttgoc tcaacttgaa cagccctatg tgttcaatca acgttcggat gotogaactg caactcattg tcatgttatg ttgagactgg tagcagaact cgaaggcatt cagtaacggtc gtagtggtga  
gacacttggt gtcccttgtcc ctcatgtggg cgaataacca gtggcttaac gcaaggttct totttgttaag aacggttaata aaggagctgg tggccatagt taaggcgcgc atctaaaagtc atttgactta ggccagcagc ttggcaactga tcttatgaa gatitttcaag aaaactggaa cactaaacac

# IDPpred Model (nastavak)

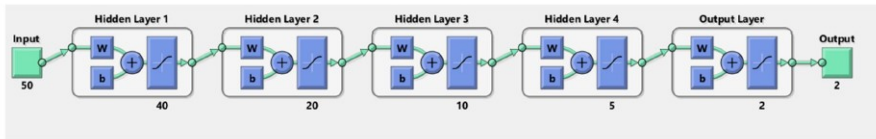
## Princip glasanja:

- 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'.

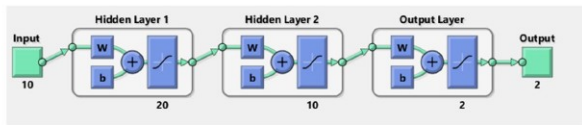
attaaagggt	tatacattcc	caggtaacaa	accaaccaac	tttcgatctc	ttgtagatct	gttctctaaa	cgaactttaa	aatctgtgtg	gtgttcaact	ggctgcatgc	ttagtgcact	cacgcagtat	aatttaatac	taattactgt	cgttgacagg	acacgagtaa	ctgtctctac	ttctgagggc
tgcttaacgt	ttgtccogtg	ttgacagcga	tcatacagac	atctaggttt	cgtccgggtg	tgaccgaaag	gtaagatgga	gagccttgtc	cctgggttca	acgagaaaaa	acacgtccaa	ctcagtttgc	ctgtttttaca	ggttcgcgac	gtgtcgtgac	gtggttttgg	agactccogtg	gaggagggtct
tatcagaggc	acgtcaacat	cttaaaagtg	gcacttgtgg	cttagtagaa	gttgaaaaag	gogttttgac	tcaacttgaa	cagccctatg	tgttcaatca	acgttcggat	gtctgaacctg	caacctcatg	tcactgttatg	gttgagctgg	tagcagaact	cgaaggcatt	cagtaacggtc	gtagtgggtga
gcacttgggt	gtccttgtcc	ctcatgtggg	cgaataacca	gtggcttaac	gcaaggttct	tctttgttaag	aacggttaata	aaggagctgg	tggccatagt	tacggcgccg	atctaaaagtc	atttgactta	ggcgacgagc	ttggcaactga	tccttatgaa	gatttttcaag	aaaaactggaa	cactaaacaa



# IDPpred Model - BPN arhitektura



(a)

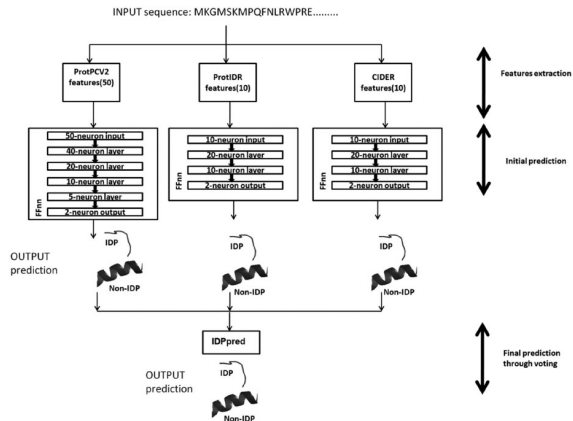


(b)

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

ataaagggtt tataccttcc caggtaacaa accaaccaac ttctgatctc ttgtagatct gtctctataa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact cagcgagtat aatttaatac taattactgt cgttgacagg acacgagtta ctgtctatct ttctgagggc  
tgcttaacgt ttgctcogtg ttgaagocga toatcagcac atctaggttt cgtccoggtg tgacocgaag gtaagatgga gagcctgtgc cctggtttca acagagaaaa ctoagtttgc atcagttocaa gttgagctgg gttgagctgg  
tatcagaggo acgtcaacet cttaagatg goacttgbtg cttagtagaa gttgaaaaag gogttttgoc toaacttgaa cagcctatg tggttcatca acgttcoggt gctgcagactg atctaaaagtc atttgactta ggccagcagc ttggcaactga tocttatgaa  
gacacttggt gtctttgtcc ctcatgtggg cgaataacca gtggcttaac gcaaggttct tocttogtaag aacggttaata aaggaagctgg tggccatagt taaggcgcgc atctaaaagtc atttgactta ggccagcagc ttggcaactga tocttatgaa  
cactaaaaaa

# IDPpred Model - Workflow



Slika: Dijagram toka rada IDPpred prediktora

attaaagggt tataccttcc caggtaacaa accaaccaac ttctgatctc ttgtagatct gtctctataa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact cagcgagtat aatttaatac taattactgt cgttgacagg acacgagtaa ctgctctatc ttctgacggc  
tgcttaacgt ttgctcogtg ttgcagocga tcatcagcac atctaggttt cgtccogggt tgacocgaag gtaagatgga gagcctgtgc cctggtttca ctcggtttca ctacgtttgc ctgctctatc gagaggtct  
tatcagaggo acgtcaacet cttaagatg gcacttggtg cttagtagaa gttgaaaaag gogttttgoc caactttgaa cagcctatg ttgttcaaca acgttoggat gctgcagactg gttgagctg tagcagaact cgaagcactt cagtaacgtc gtagtggtga  
gcacttggt gtctttgtcc ctcatgtggg cgaataacca gtggtttacc gcaaggttct ttctgttaag aacgttaata acgttgctg ttgcatagt taaggcgcgc atctaaagtc atttgactta ggccagcagc ttgccaactg tcttatgaa gattttcaag aaaaactgaa cactaaacac

## 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):

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

attaaagggt tattaotttcc caggtaacaa accaaccaac ttctgatctc ttgtagatct gtctctataa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact caccagttat aatttaatac taattactgt cgttgacagg acacagttaa ctgtctatc ttctgaggc  
tgcttaacgt ttgtccogtg ttgaagocga tctatcagac atctaggttt cgtccogggt tgacgcgaag gtaagatgga gagcctgtgc cctgggttca acacgtccaa ctoagtttgc ctgttttaca ggttcogac gtgctcgtac gaggaggtc  
tatcagaggo acgtcaacct cttaagatg gaacttgttg cttagtagaa gttgaaaaag gogttttgoc tcaacttgaa cagcctatg tgttcaatca acgttcoggt gctcgaactg cgttgagctg tagcagaact cgaaggcatt cagtaocgtc gtagtggtg  
gacacttggt gtctttgtcc ctoatgtggg cgaataacca gtggcttaac gcaaggttct totttogtaag aacggttaata aaggagctgg tggccatagt taaggcgccg atctaaaagtc atttgactta ggccacgagc ttggcaactg tcttatgaa gattttcaag aaaaactgaa cactaaacac

## IDPpred Evaluacija (nastavak)

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

**F-mjera (F-measure):**

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

**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)

$N_{\text{dis}}$ : ukupan broj nestrukturiranih proteina

$N_{\text{ord}}$ : ukupan broj strukturiranih proteina

attaaagggt	tataccttcc	caggtaacaa	acaaacaaac	tttcgatctc	ttgtagatct	gtctctataa	cgaactttaa	aactctgtgt	gtgtcactc	ggctgcatgc	ttagtgcact	cacgcagtat	aatttaatac	taattactgt	cgttgacagg	acacgagtaa	ctcgtctatc	ttctgaggc
tgcttaacgt	tttgtccgtg	ttgaagocga	tcatacgaac	atctaggttt	cgtccgggtg	tgacocgaag	gtaagatgga	gagcctgtgc	ctcgtgttca	acgagaaaaa	acacgtccaa	ctcagtttgc	ctgtttttaca	ggttcgcgac	gtgtcgtgac	gtgcttttgc	agactccgtg	gaggaggtct
tatcagaggo	acgtcaaacct	cttaaaagtg	gcacttgttg	cttagtagaa	gttgaaaaag	gggtttttgc	tcaacttgaa	cagccctatg	tgttcaatca	acgttcggat	gotcgaactg	caactccatg	tcactgttatg	gttgagctgg	tagcagaact	cgaaggcaat	cagtaacggtc	gtagtgggtg
gcacttgggt	gtccttgtcc	ctcatgtggg	cgaataacca	gtggcttaac	gcaaggttct	tctttgttaag	aacggttaata	aaggagctgg	tggccatagt	tacggcgcgg	atctaaaagtc	atttgactta	ggcgcagcgc	ttggcaactga	tccttatgaa	gattttcaag	aaaaactgaa	cactaaacac

## IDPpred Rezultati

Predictors	Performance evaluation parameters*									
	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
fIDPnn	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

attaaagggt tataccttcc caggtaacaa accaaccaac ttctgatctc ttgtagatct gtctctctaa cgaactttaa aatctgtgtg gctgtcactc gctgtcactc ttagtccact cagccagtat aatttaatac taattactgt cgttgacagg acacagtaga ctgtctctac ttctgcaggc  
 tgcctaacgt ttgtccogtg ttgcagcaga tcatcagcac atctaggttt cgtccgggtg tgaccgaaag gtaagatgga gagcctgtgc cctggtttca acacgtccaa ctacgttttg ctgtttttac ggtttgcgac gtgctcgtac gtgcttttg  
 tatcagaggo acgtcaacat cttaaaagtg gaacttgttg cttagtagaa gttgaaaaag gogtttttgc tcaacttgaa cagccctatg tgttcaatca acgttcoggt gctcgaactg caactcattg tcatgtttat gttgagctgg tagcagaact cgaagcgact  
 gacacttggt gtccttgtcc ctcatgtggg cgaataacca gtggcttaac gcaaggtttc tcttctgaag aacgttaata aaggaactgg tggccatagt taaggcgcgg atctaaaagtc atttgactta ggcgcagcgc ttggcaactga tcttatgaa gattttcaag aaaaactgaa cactaaacac

# Zaključak

- IDPpred je novi i efikasn 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

attaaggtt	tataccttcc	caggttaacaa	acaaacaaac	tttogatctc	ttgtagatct	gttctctaaa	cgaactttta	aattctgtgtg	gctgtcaactc	ggctgcactgc	ttagtgcact	cacgcagtat	aattaataac	taattactgt	cggtgacagg	acaagagtaa	ctcgtctatc	ttctgaggc
tgctaacggt	ttcgtccgtg	ttgcagccga	tcctcagcac	atctaggttt	cgtccgggtg	tgacccgaaag	gtaagatgga	gagccttgtc	cctggtttca	acgagaaaaac	acacgtccaa	ctcagtttgc	ctgtttttaca	ggttccggac	gtgctcgtac	gtggctttgg	agactccgtg	gaggaggtct
tatcagaggo	acgtcaacat	cttaagatg	gcacttggtg	cttagtagaa	gttgaaaaag	gogttttgoc	tcaacttgaa	cagccctatg	tggtccatcaa	acgttcggat	gctcgaaactg	cacctccatg	tcctgttatg	gttgagctgg	tagcagaact	cgaaggcaatt	cagtaacggtc	gtagtgggtga
gacacttggt	gtccttgtoo	ctcatgtggg	cgaataacca	gtggottacc	gcaaggttct	tcttctgtaag	aacggtaata	aaggagctgg	tggccatagt	taaggcgccg	atctaaagtc	atttgactta	ggcgacgago	ttggcaactga	tccttatgaa	gattttcaag	aaaactgaaa	cactaaacac

# 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 fIDPnn2 - Wang et al. (May 2024)

attaaggtt tataccttcc caggttaaca accaaccac tttagatctc ttgtagatct gttctotaaa cgaactttta aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact caccgagtat aattaataac taattactgt cgttgacagg acacagagtaa ctgctctatc ttctgaggc  
tgctaacggt ttgctcogtg ttgcagcoga tcatcagcac atctaggttt cgtccgggtg tgacogaaag gtaagatgga gagccttgtc cctggtttca acagagaaaac acacgtocaa ctacagtttg ggttgcogac gtgctcgtac gtggttttg agactcogtg gaggaggtct  
tatcagaggo acgtcaacat cttaaagatg gcacttgtyg cttagtagaa gttgaaaaag gogttttgoc tcaacttgaa cagcctatg tgttcatcaa acgttcggat gctogaactg atctaaagtc atttgactta ggccagcago ttggcaactg tcttatgaa gattttcaag aaaactgaa cactaaace  
gacacttggt gtccctgtcc ctcactgtgg cgaataacca gtggtttacc gcaaggttct tottctgaag aacggttaata aaggagctgg tggccatagt taaggcgcgc atctaaagtc atttgactta ggccagcago ttggcaactg tcttatgaa gattttcaag aaaactgaa cactaaace

# 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*.

attaaggtt	tataccttc	caggtaacaa	accaaccaac	tttogatctc	ttgtagatct	gttctotaaa	cgaactttaa	aatctgtgtg	gtgtgcactc	ggctgcatgc	ttagtgcact	cacgcagtat	aattaataac	taattactgt	cgttgacagg	acaagagtaa	ctgtctatct	ttctgagggc
tgctaacggt	tttgtccgtg	ttgcagccga	tcctcagcac	atctaggttt	cgtccgggtg	tgacccgaag	gtaagatgga	gagccttgtc	cctgggttca	acgagaaaaac	acaagtcocaa	ctcagtttgc	ctgtttttaca	ggttcgcgac	gtgctcgtac	gtggctttgg	agactccgtg	gaggaggtct
tatcagaggc	acgtcaacat	cttaaaagatg	gcacttgttg	cttagtagaa	gttgaaaaag	gggtttttgoc	tcaacttgaa	cagccttatg	tgttcatcaa	acgttcggat	gtctogaactg	cacctccttg	tcattgttatg	gttgagctgg	tagcagaact	cgaaggcaatt	cagtacggctc	gtagtggtag
gacacttggt	gtccttgtcc	ctcatgtggg	cgaataacca	gtggottacc	gcaaggttct	tcttctgtaag	aacggttaata	aaggagctgg	tggccatagt	taaggcgccg	atctaaagtc	atttgactta	ggcgacgagc	ttggcaactga	tccttatgaa	gattttcaag	aaaactgcaa	cactaaacac