Predviđanje strukture proteina modelom AlphaFold 2

Jelušić, Darijan

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
Preddiplomski sveučilišni studij računarstva

Završni rad

Predviđanje strukture proteina modelom AlphaFold 2

Rijeka, rujan 2022.

Darijan Jelušić
0069087847

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
Preddiplomski sveučilišni studij računarstva

Završni rad

Predviđanje strukture proteina modelom AlphaFold 2

Mentor: doc. dr. sc. Goran Mauša

Rijeka, rujan 2022.

Darijan Jelušić
0069087847

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
POVJERENSTVO ZA ZAVRŠNE ISPITE

Rijeka, 14. ožujka 2022.

Zavod: **Zavod za računarstvo**
Predmet: **Uvod u objektno orijentirano programiranje**
Grana: **2.09.04 umjetna inteligencija**

ZADATAK ZA ZAVRŠNI RAD

Pristupnik: **Darijan Jelušić (0069087847)**
Studij: Preddiplomski sveučilišni studij računarstva

Zadatak: **Predviđanje strukture proteina modelom AlphaFold 2 / Protein structure prediction using AlphaFold 2 model**

Opis zadatka:

Proučiti arhitekturu i sastavne module programskog modela AlpaFold 2 koji je razvijen za predviđanje 3D strukture proteina na osnovu kompozicije amino kiselina. Objasniti primjenu koncepata evolucijskog i strukturnog pretraživanja, tehnike sastavljanja i iterativnog rafiniranja predviđene strukture te metrike za procjenu pouzdanosti završne strukture. Istražiti mogućnost primjene modela AlphaFold 2 za predviđanje strukture kraćih peptidnih sekvenci.

Rad mora biti napisan prema Uputama za pisanje diplomskih / završnih radova koje su objavljene na mrežnim stranicama studija.

Zadatak uručen pristupniku: 14. ožujka 2022.

Mentor:

Doc. Goran Mauša, dipl. ing.

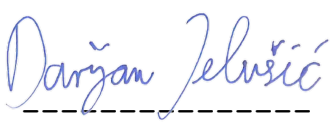
Predsjednik povjerenstva za završni ispit:

Prof. dr. sc. Kristijan Lenac

Izjava o samostalnoj izradi rada

Izjavljujem da sam samostalno izradio ovaj rad.

Rijeka, rujan 2022.



Darijan Jelušić

Zahvala

Zahvaljujem svom mentoru doc. dr. sc. Goranu Mauši na pomoći i smjernicama tijekom izrade ovog rada te cijelom DeShPet timu na korisnim informacijama. Zahvaljujem i obitelji i prijateljima na pruženoj podršci i razumijevanju tijekom studiranja.

Zahvaljujem Sveučilišnom računskom centru Sveučilišta u Zagrebu čiji su resursi korišteni u izradi ovog rada.

Sadržaj

Popis slika

Popis tablica

Poglavlje 1

Uvod

Ovaj završni rad dio je uspostavnog istraživačkog projekta Hrvatske zaklade za znanost pod naslovom "Dizajn katalitički aktivnih peptida i peptidnih nanostruktura", s oznakom UIP-2019-04-7999. Cilj rada bio je proučavanje arhitekture i sastavnih modula programskog modela AlphaFold 2, objašnjavanje koncepata primijenjenih u njegovoj izgradnji te istraživanje mogućnosti primjene modela za predviđanje trodimenzionalne strukture kraćih peptidnih sekvenci.

1.1 Problem savijanja proteina

Proteini, odnosno bjelančevine, obnašaju širok spektar funkcija u svijetu biokemije, te su jedna od najznačajnijih tvari u svim živim bićima [1]. Po svojoj građi, proteini su dugi lanci aminokiselina povezanih peptidnim vezama, odnosno polipeptidi - peptidi s barem pedesetak aminokiselina u lancu [2]. Funkcija peptida definirana je njegovom tercijarnom (trodimenzionalnom) strukturom, koja je pak definirana samim nizom aminokiselina od kojih se sastoji lanac tog peptida [1, 3]. Tercijarna struktura peptida nije jednostavna, lanac aminokiselina nije ispružen, već se prirodno savija u kompleksne strukture [4]. Upravo u tome leži problem savijanja proteina, koji je nastao sredinom dvadesetog stoljeća, kada su Kendrew i suradnici eksperimentalno odredili tercijarnu strukturu mioglobina i ostali zaprepašteni njenom kompleksnošću te nedostatkom pravilnosti i simetrije [1, 5]. To je postavilo tri

ključna pitanja glede trodimenzionalne strukture proteina i peptida: Na koji način jednodimenzionalan niz aminokiselina utječe na konačnu trodimenzionalnu strukturu, koji je mehanizam samog savijanja te je li moguće računalno predvidjeti trodimenzionalne strukture proteina na temelju njegove sekvence aminokiselina [6]. U svrhu pronalaska odgovora na posljednje pitanje, 1994. godine pokrenuto je natjecanje u računalnom predviđanju tercijarnih struktura proteina *Critical Assessment of protein Structure Prediction* - CASP. Natjecanje se održava svake dvije godine, i cilj mu je unapređivanje i testiranje modela za predviđanje strukture proteina iz sekvenci aminokiselina [7].

1.2 AlphaFold 2

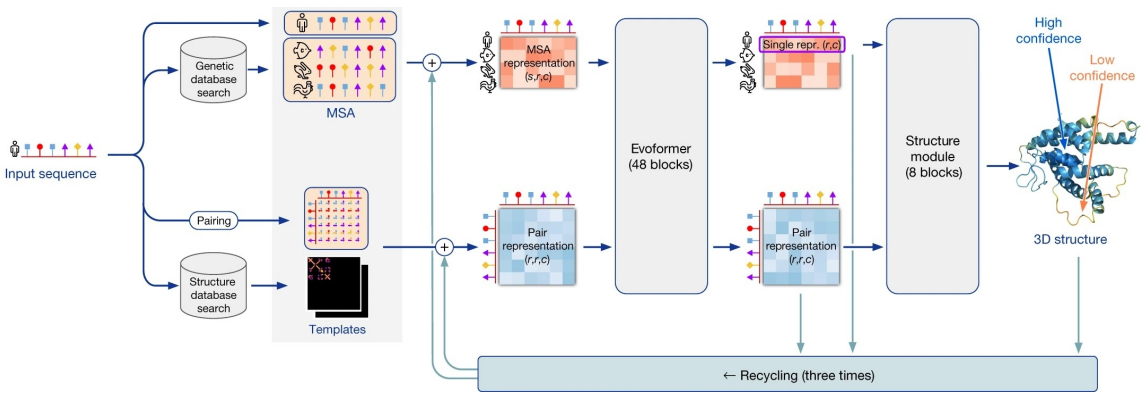
Na četrnaestom CASP natjecanju 2020. godine Googleov DeepMind tim predstavio je drugu iteraciju svog modela za predviđanje strukture proteina - AlphaFold 2 [8]. Taj je model pokazao iznimnu točnost i konzistentnost u predviđanju tercijarne strukture proteina, nerijetko bivajući na razini točnosti eksperimentalno određenih struktura, čak i u kategoriji *ab initio* modeliranja, odnosno modeliranja proteina koji nemaju *puno* sličnosti s poznatim strukturama [9]. AlphaFold 2 ima i svoja ograničenja, značajno se oslanja na količinu srodnih proteina, nema mogućnost predviđanja post-translacijskih modifikacija i nekanonskih aminokiselina, te je značajno ograničen glede predviđanja intrinzično neuređenih struktura i same strukturne dinamike proteina [10, 11]. Usprkos tome, mogućnosti AlphaFold 2 modela mnogi smatraju važnima u kontekstu rješavanju problema savijanja proteina [12].

Poglavlje 2

Pregled AlphaFold 2 i važnih pojmova

2.1 AlphaFold 2

AlphaFold 2 mreža arhitekturalno je razložena na tri glavne cjeline: Pretraživanje, Evolucijski transformator (*Evoformer*) i Strukturni modul [8, 13]. Grafički prikaz same strukture vidljiv je na Slici 2.1.



Slika 2.1 Pojednostavljeni prikaz strukture AlphaFold 2 modela, preuzeto iz [8]

Tok podataka kroz mrežu je sljedeći:

1. Protein čija se struktura želi odrediti preda se kao lanac aminokiselina predstavljen u formatu FASTA [14].
2. Pretražuju se genetske baze proteina kako bi se pronašle sekvence aminokiselina slične ulaznoj, te se iz njih gradi višestruko poravnanje sekvenci - Multiple Sequence Alignment (MSA).
3. Uz pomoć izrađenog MSA, pretražuju se baze struktura proteina kako bi se pronašao strukturni predložak kojim će se izgraditi reprezentacija parovima (eng. *pair representation*) [8].
4. Dobivena MSA reprezentacija i reprezentacija parovima pretprocesiraju se i međusobno nadopunjavaju prije nego što se pošalju dalje u *Evoformer*.
5. *Evoformer* iterativno rafinira i nadograđuje reprezentaciju MSA i reprezentaciju parovima te izmjenjuje podatke između njih kako bi ulazni podaci strukturniog modula bili što informiraniji i ispravniji. [13].
6. Strukturni modul iz reprezentacije MSA i reprezentacije parovima koje je dobio iz *Evoformera* gradi trodimenzionalni model strukture predstavljen nizom 3D Kartezijevih koordinata svih atoma u lancu.

Izlazi strukturnog modula (koordinate atoma, nadograđene reprezentacija MSA i reprezentacija parovima i dodatni izlazi) dodatno se tri puta šalju na ulaz *Evoformera* kako bi se iterativno proizvela što preciznija konačna struktura.

2.2 Višestruko poravnanje sekvenci

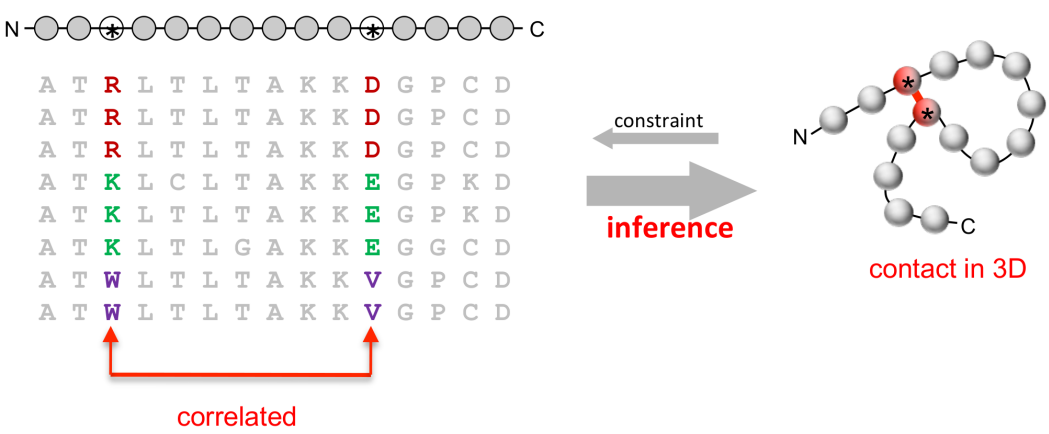
Peptidi i proteini po svojoj su građi lanci aminokiselina povezanih peptidnim vezama. Iako je otkriveno preko petsto različitih aminokiselina, od kojih se sto četrdeset nalazi u prirodnim proteinima, tih sto četrdeset nastaje post-translacijskim modifikacijama samo 22 temeljne aminokiseline koje se pronalaze u genetskim uputama svih živih bića [15]. Za računalni zapis i obradu proteina, najčešće se koristi FASTA [14] zapis u kojem se svaka od 22 temeljne aminokiseline predstavlja jednim velikim slovom engleske abecede uz tri posebna znaka za nepoznate aminokiseline u

lancu i završetak translacije [16]. Takve sekvence aminokiselina mogu se uspoređivati s ciljem pronalaženja sličnosti te promatranja evolucije i evolucijske povezanosti različitih proteina. Kako bi se olakšalo uspoređivanje, najčešće se koristi tehnika višestrukog poravnanja sekvenci (eng. Multiple Sequence Alignment (MSA)), gdje se nekoliko sekvenci „poravna" ovisno o sličnim regijama i zajedničkim obilježjima. Primjer poravnanja nekoliko sekvenci nukleinskih kiselina vidljiv je na Slici 2.2.

Sequence1	**-TCAGGA-TGAAC----**
Sequence2	**ATCACGA-TGAACC---**
Sequence3	**ATCAGGAATGAATCC--**
Sequence4	**-TCACGATTGAATCGC-**
Sequence5	**-TCAGGAATGAATCGCM**

Slika 2.2 Poravnanje pet sekvenci nukleinskih kiselina, preuzeto iz [17]

Kod predviđanja trodimenzionalne strukture proteina, višestruko poravnanje sekvenci može pomoći pri određivanju koje su rezidue u lancu nekog proteina fizički blizu. Uzevši u obzir da su evolucijski srodni proteini slični po funkciji i strukturi, te da je sama funkcija nekog lanca aminokiselina definirana njegovom strukturom, korisno je prilikom izračuna strukture proteina uzeti u obzir njegove evolucijske srodnike [3]. Unutar MSA evolucijski srodnih proteina, može se pratiti koji su se parovi rezidua koje nisu susjedne u lancu „zajedno" mijenjale između različitih proteina, te iz toga zaključiti koji su parovi fizički povezani u tercijarnoj (trodimenzionalnoj) strukturi proteina [18].



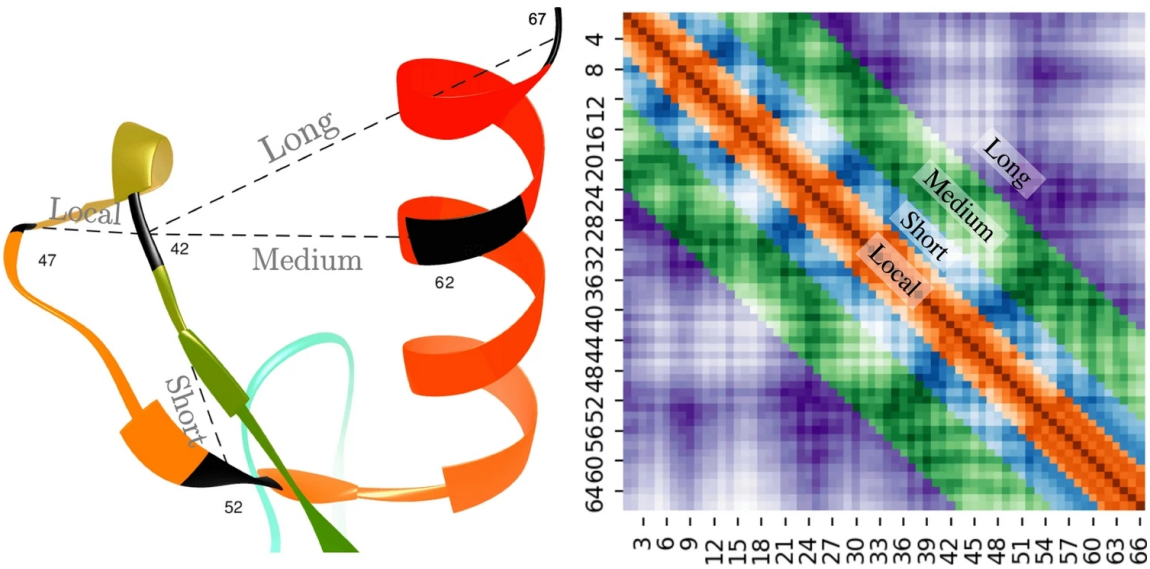
Slika 2.3 Određivanje povezanosti rezidua u tercijarnoj strukturi proteina temeljem koevolucije aminokiselina unutar lanca, preuzeto iz [18]

Na Slici 2.3 prikazana je teoretska primjena tog koncepta u modelu za predviđanje tercijarne strukture proteina na temelju lanca aminokiselina. U MSA na lijevoj strani slike, vidljivo je da se rezidue na trećem i jedanaestom mjestu u lancu uvijek zajednički mijenjaju između različitih sekvenci u tom poravnanju. Temeljem toga, može se zaključiti da su te dvije pozicije u lancu fizički povezane, te se time može informirati izgradnju tercijarne strukture. Ujedno se pretpostavkom da su te dvije rezidue povezane mogu ograničiti pozicije rezidua susjedne onima iz povezanog para, čime se povratno informira istraživanje MSA te daljnji zaključci koji se iz njega izvode.

2.3 Distogram

Koristan resurs prilikom izračuna tercijarne strukture proteina jest i takozvani *distogram* odnosno matrica udaljenosti rezidua lanca. Distogram proteina u suštini je matrica kojoj su i redci i stupci indeksi rezidua unutar lanca koji sačinjava taj protein. Vrijednost u matrici na mjestu (i, j) kazuje euklidsku udaljenost u tercijarnoj strukturi između i-te i j-te rezidue lanca [19]. Stoga su vrijednosti svih udaljenosti na glavnoj dijagonali $(i = j)$ jednake nuli. Uz matricu udaljenosti, nerijetko se koristi i matrica kontakta, koja prikazuje koje su rezidue u međusobnom kontaktu. Definicija kontakta ovisi o kontekstu, ali u većini slučajeva se za dvije rezidue smatraju da su

u kontaktu ako su udaljene manje od neke zadane vrijednosti, npr. 8 Å ili 12 Å [1] [19].



Slika 2.4 Prikaz odnosa udaljenosti rezidua na lancu i udaljenosti od glavne dijagonale u matrici udaljenosti, preuzeto iz [19]

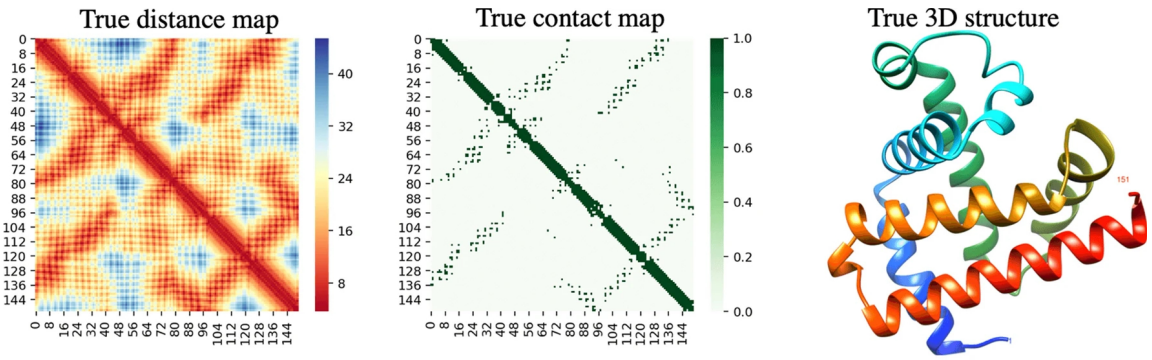
Tablica 2.1 Definicije regija distograma

Regija	Broj rezidua razmaka
Lokalna	< 5
Kratka	6 - 11
Srednja	11 - 23
Duga	> 23

Za kvalitetno predviđanje strukture proteina, bitno je, osim same fizičke udaljenosti, uzeti u obzir i koliko su rezidue međusobno udaljene u lancu. Što je pozicija u matrici udaljenosti više odmaknuta od glavne dijagonale, to su te rezidue udaljenije na lancu aminokiselina. Na Slici 2.4 prikazane su proizvoljno odabrane zone oko glavne dijagonale unutar kojih se za dvije rezidue smatra da su međusobno lokalne (*Local*) ili na kratkoj (*Short*), srednjoj (*Medium*) ili dugoj (*Long*) udaljenosti. Na

[1] Angstrem, $1Å = 10^{-10}m$

ovom konkretnom primjeru, brojevi rezidua razmaka koje određuju te zone definirane su u Tablici 2.1 [19]. Na raznim se primjerima pokazalo da su za veću točnost predviđene strukture proteina od veće važnosti udaljenosti između rezidua koje su na duljoj udaljenosti unutar lanca, jer više govore o samom savijanju lanca [20].



Slika 2.5 Matrice udaljenosti i kontakta te 3D struktura proteina PDB 1A6M, preuzeto iz [19]

Na Slici 2.5 prikazane su matrica udaljenosti (*True distance map*) i kontakta (*True contact map*) te 3D prikaz tercijarne strukture proteina (*True 3D structure*). Vrijednosti unutar matrica radi zornosti su prikazane bojama umjesto brojevima. U slučaju matrice udaljenosti, vrijednosti u angstremima prikazane su bojama između crvene i plave, gdje crvena predstavlja 0 Å, a plava 48 Å. Vrijednosti u matrici kontakta zapravo su binarne, odnosno ispunjeno polje predstavlja 1 a prazno 0, gdje 1 označava da rezidue jesu u kontaktu a 0 da nisu.

Poglavlje 3

Pretraživanje

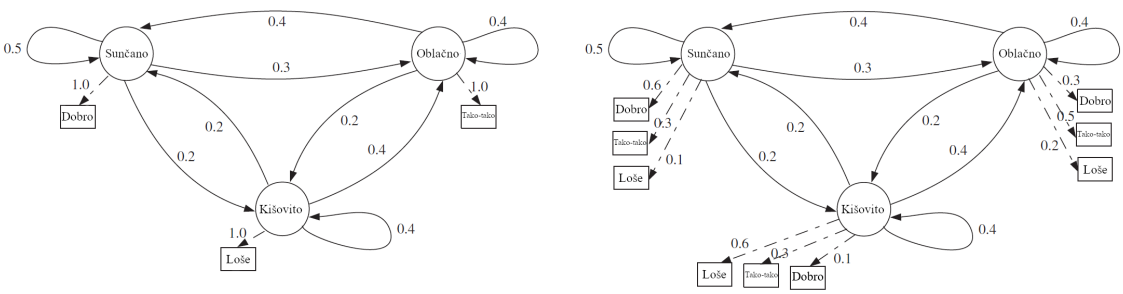
Proces predviđanja tercijarne strukture proteina u mreži AlphaFold 2 započinje pretraživanjem baza proteina kako bi se pronašli lanci aminokiselina slični i srodni onom čija se struktura želi odrediti, te na temelju kojih se grade reprezentacija MSA i reprezentacija parovima. AlphaFold 2 u tu svrhu koristi već ustanovljene aplikacije za pretragu baza: JackHMMER, HHBlits i HHSearch [21–23]. Sva tri alata za pretragu temelje se na upotrebi skrivenih Markovljevih modela (eng. Hidden Markov Model (HMM)) u svrhu pronalaženja homologije između proteinskih sekvenci [24].

3.1 Skriveni Markovljev model

Skriveni Markovljev model vjerojatnosni je model kojim se opisuju nizovi opažanja Markovljevog procesa koji se ne može direktno promatrati [25, 26]. Markovljev proces, odnosno Markovljev lanac, način je modeliranja stohastičkog procesa s diskretnim stanjima uz pretpostavku da svako novo stanje u koje sustav prelazi ovisi isključivo o stanju koje mu prethodi, odnosno da zadovoljava tzv. Markovljevo svojstvo, formalno zapisano izrazom 3.1:

$$P(X_n = x \mid X_{0:n-1}) = P(X_n = x \mid X_{n-1}) \tag{3.1}$$

gdje X_n predstavlja n-to stanje Markovljevog lanca, a $X_{0:n-1}$ sva stanja koja su se u lancu našla prije n-tog stanja, gdje su stanja $x_i \in S$ gdje S simbolizira *prostor stanja* tog Markovljevog lanca, skup koji sadrži sva moguća stanja tog sustava. [25, 27]. Tada $P(X_n = x \mid X_{0:n-1})$ predstavlja vjerojatnost da n-to stanje sustava bude x uzevši u obzir sva stanja koja su prethodila n-tom, što se još može zapisati i kao $P(X_n = x \mid X_0 = x_0, X_1 = x_1, ..., X_{n-1} = x_{n-1})$. Markovljevo svojstvo govori da je stanje X_n nezavisno o stanjima koja prethode stanju $n-1$ uz uvjet da je poznato stanje X_{n-1}, odnosno formalnim zapisom: $X_n \perp\!\!\!\perp X_{0:n-2} \mid X_{n-1}$. Ključno svojstvo skrivenog Markovljevog modela po kojem se razlikuje od običnog jest da stanja lanca, kao ni procese prijelaza između stanja, nije moguće promatrati, već su uočljiva samo *opažanja* koja nisu jednoznačno povezana sa samim stanjima [28].



(a) Markovljev lanac s opažanjima *(b) Skriveni Markovljev lanac s opažanjima*

Slika 3.1 Dijagrami stanja i prijelaza običnog i skrivenog Markovljevog lanaca s opažanjima

Razlika između običnog i skrivenog Markovljevog lanca vidljiva je na Slici 3.1 [28] gdje su za obje vrste Markovljevog modela prikazani dijagrami stanja i prijelaza, te odnos opažanja i stanja. Oba primjera opisuju vremensku prognozu i raspoloženja (opažanja) vezana uz određena stanja prognoze. Krugovi predstavljaju stanja, pravokutnici opažena raspoloženja, pune strelice i pripadne vrijednosti prikazuju moguće prijelaze između stanja i vjerojatnosti samih prijelaza, a isprekidane strelice i pripadne vrijednosti pokazuju uzročnost stanja i opažanja uz pripadne vjerojatnosti.

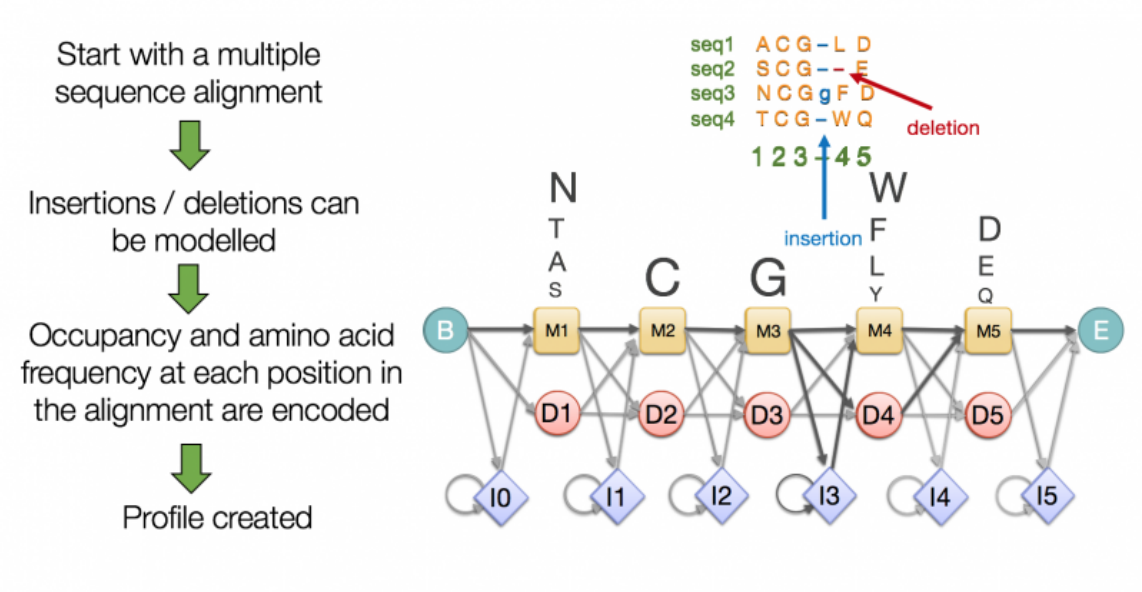
Prostor stanja sadrži tri moguće vremenska prognoze: sunčano, oblačno i kišovito. Raspoloženja koja mogu biti opažena su: dobro, loše i tako-tako. Na Podslici 3.1a vidljivo je da svaka vremenska prognoza uvijek rezultira istim opaženim raspoloženjem, npr. ako je vrijeme *sunčano*, opaženo raspoloženje je *dobro* u 100% slučajeva, pa se može zaključiti i suprotno - ako je raspoloženje *dobro*, može se sa sigurnošću utvrditi da je vrijeme *sunčano*. To je primjer običnog Markovljevog lanca, jer se stanja, iako indirektno, mogu opažati. To nije slučaj kod skrivenih Markovljevih lanaca, kao što je onaj prikazan Podslikom 3.1b. Vidljivo je da svako stanje može, s određenom vjerojatnošću, uzrokovati svako raspoloženje. Ako je vrijeme, kao u prethodnom primjeru, *sunčano*, iz dijagrama se može očitati da će opaženo raspoloženje u 60% slučajeva biti *dobro*, u 30% *tako-tako* te u preostalih 10% *loše*. Stoga se iz opažanja ne može jednoznačno odrediti stanje lanca, te se takvi modeli nazivaju *skrivenim* Markovljevim modelima [28].

Profilni skriveni Markovljev model

Poseban slučaj skrivenih Markovljevih modela, koristan u kontekstu bioloških sekvenci, upotrebljava se za modeliranje MSA. Model u pitanju je takozvani *profilni* skriveni Markovljev model - Profile Hidden Markov Model (PHMM) [26, 29]. Nastao je 1994. godine upravo u kontekstu problematike modeliranja proteina, te je PHMM zapravo primjena koncepta skrivenog Markovljevog modela u svrhu statističkog modeliranja, pretraživanja baza te izrade MSA proteinskih obitelji i domena [26]. Kako se prostor stanja u svijetu proteina sastoji od dvadeset poznatih proteinogenih aminokiselina, Markovljevi lanci vrlo su zahvalan temelj za modeliranje proteina [26, 28]. Specifičnost PHMM-a je što su definirana tri vrste stanja: podudaranje (*match*), umetanje (*insert*) i brisanje (*delete*) [26].

Na Slici 3.2 prikazan je dijagram stanja i prijelaza PHMM-a, gdje su stanja podudaranja prikazana slovom M u žutom kvadratu, stanja umetanja slovom I u plavom rombu te stanja brisanja slovom D u crvenom krugu.

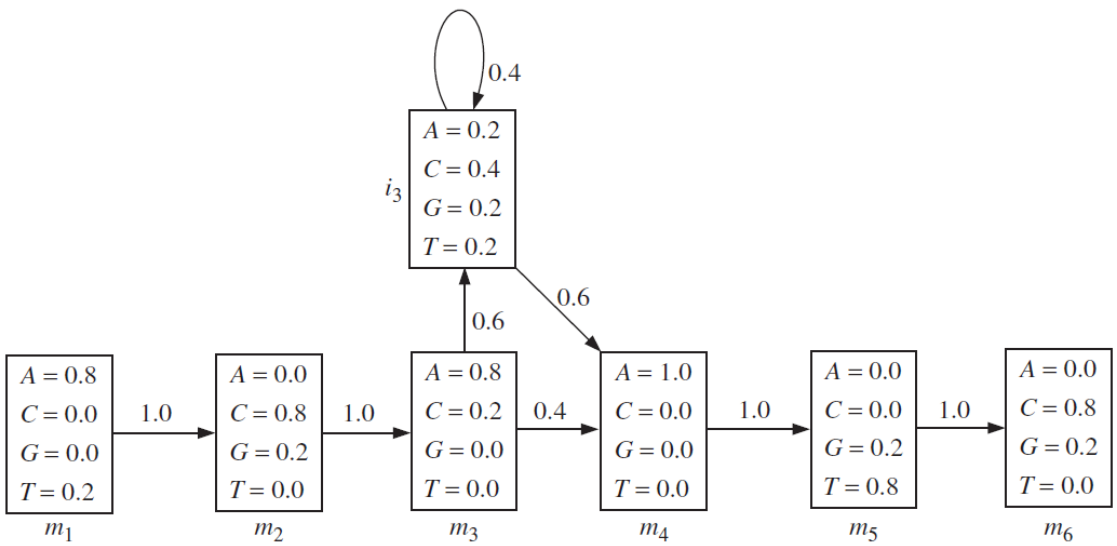
Stanja podudaranja - M stanja - kazuju koliko je vjerojatno da će se na nekoj poziciji u sekvenci naći određena aminokiselina, i svako stanje podudaranja ima svoju razdiobu vjerojatnosti za 20 mogućih aminokiselina koje se mogu naći na tom mjestu



Slika 3.2 Dijagram stanja i prijelaza PHMM-a, preuzeto iz [30]

[26]. Pojedino stanje podudaranja označava određenu poziciju u lancu aminokiselina, te ako se protein sastoji od n aminokiselina, PHMM sadrži n stanja podudaranja ($n = 5$ za primjer na Slici 3.2). Sažeto rečeno, svako od n stanja podudaranja m generira aminokiselinu x na mjestu i s vjerojatnošću $P(x \mid m_i),\ i = 1...n$. Za svako od n stanja podudaranja, postoji stanje brisanja, koje je zapravo *„dummy"* stanje kojim se preskače to stanje podudaranja, odnosno ubaci praznina umjesto jedne od 20 aminokiselina. Naposljetku, na n stanja podudaranja postoji $n+1$ stanja umetanja - po jedno sa svake strane stanja podudaranja. Stanja umetanja funkcioniraju slično stanjima podudaranja, jer svako od $n+1$ stanja umetanja i_k generira aminokiselinu x s vjerojatnošću $P(x \mid i_k),\ k = 0...n$. Iz svakog stanja moguća su tri prijelaza (osim stanja brisanja iz kojeg se ne može prijeći u stanje umetanja), prikazana strelicama na Slici 3.2. Prijelazi u stanja podudaranja ili brisanja uvijek napreduju kroz model, dok stanja umetanja mogu prelaziti sama u sebe. Vjerojatnost prijelaza između dva stanja p i q iznosi $T(q \mid p)$. Sekvenca se iz PHMM-a može generirati „nasumičnom šetnjom" kroz model, krećući iz početnog stanja modela (B na primjeru sa Slike 3.2), prelazeći pritom u iduće stanje m_1, i_1 ili d_1 s vjerojatnostima $T(m_1 \mid B)$, $T(i_1 \mid B)$ i $T(d_1 \mid B)$, i tako napredujući kroz model sve dok se ne dođe do završnog stanja (E na primjeru sa Slike 3.2), čime je generirana jedna moguća sekvenca [26].

Stvaranje „profila" nekog MSA pomoću PHMM-a svodi se na određivanje matrice emisijskih vjerojatnosti i matrice prijelaza. Matrica emisijskih vjerojatnosti sadrži vjerojatnosti da će se određeni dokaz (u kontekstu proteina, određena aminokiselina) opaziti ovisno o stanju Markovljevog modela, što u ovom slučaju predstavlja vjerojatnost da će se svaka od aminokiselina pojaviti u određenom stupcu MSA. Matrica prijelaza sadrži vjerojatnosti prijelaza između različitih stanja Markovljevog lanca, odnosno vjerojatnosti da prijeđe iz stanja p u stanje q za svaki p i q modela ($T(q \mid p)$), što se najjednostavnije može postići prebrojavanjem svih prijelaza [28].



Slika 3.3 Primjer PHMM-a za molekule DNK, preuzeto iz [28]

Kako se moguće sekvence lanca stvaraju „nasumičnom šetnjom" kroz Markovljev lanac, tako se i svaka sekvenca iz MSA može predstaviti jednom šetnjom kroz Markovljev lanac. Vjerojatnost svake sekvence može se izračunati množenjem matrica emisijskih vjerojatnosti i matrica prijelaza po putu te sekvence. Slika 3.3 sadrži primjer PHMM-a za molekule DNK, gdje su opažanja dušične baze (A, C, G, T). U svakom m i i stanju nalaze se vjerojatnosti da će se na tom mjestu opaziti određena dušična baza, a na strelicama između stanja nalaze se vjerojatnosti da će se dogoditi upravo taj prijelaz. Za sekvencu $AGCATG$ izračun vjerojatnosti bio bi sljedeći:

$$0,8 * 1,0 * 0,2 * 1,0 * 0,8 * 0,4 * 1,0 * 1,0 * 0,8 * 1,0 * 0,2 = 0,008192 \qquad (3.2)$$

Budući da su vjerojatnosti nužno manje od 1 i lako može doći do *underflowa* i gubitka preciznosti, redovito se umjesto samog množenja računa logaritam te vjerojatnosti, pa bi logaritamska vrijednost za tu sekvencu bila:

$$3ln(0,8) + 5ln(1) + 2ln(0,2) + ln(0,4) = -4,8046 \tag{3.3}$$

gdje ln predstavlja prirodni logaritam, odnosno logaritam po bazi e. Ovakve su vrijednosti manje računski zahtjevne, te veća vrijednost predstavlja veću vjerojatnost da je sekvenca dio profiliranog MSA [28].

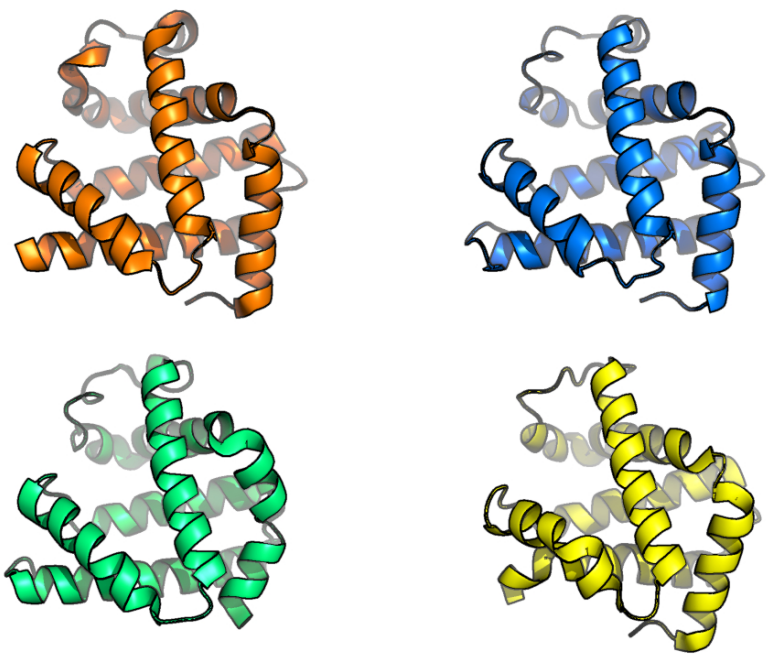
3.2 Genetsko pretraživanje

Cilj genetskog pretraživanja jest izgradnja što podrobnijeg MSA za ulazni protein, tražeći srodne proteine u genetskim bazama - najčešće proteine koji vrše iste funkcije u drugim organizmima. Za pretragu se koriste alati JackHMMER v3.3 [21] i HHBlits v3.0-beta.3 [22], te pretražuju genetske baze MGnify [31], UniRef90 [32], Uniclust30 [33] i BFD [34] [8]. Konkretno se alatom JackHMMER pretražuju MGnify i UniRef90, a HHBlits Uniclust30 i BFD. Dubina MSA ograničena je na pet tisuća sekvenci za MGnify, deset tisuća za UniRef90, a neograničena za sve pretrage alatom HHBlits. Iz dobivenih MSA uklone se duplikati te se ujedine u jedan MSA za daljnje procesiranje, gdje je prvi red matrice upravo sekvenca ulaznog proteina. Stupci MSA predstavljaju lokacije u lancu proteina, a redci različite sekvence [8].

3.3 Strukturno pretraživanje

Kako bi mreža što informiranije zaključivala trodimenzionalnu strukturu proteina, stvara se reprezentacija parovima koja se gradi temeljem strukturnih predložaka. Nakon pretrage genetskih baza, MSA dobiven pretraživanjem UniRef90 baze koristi se kao ulaz za HHSearch alat [23] koji traži strukturne predloške u PDB70 [35] strukturnoj bazi. Strukturni podaci iz svakog podudaranja ekstrahiraju se iz mmCIF [36] datoteka iz PDB baze [8]. U slučaju da se sekvenca iz PDB70 ne podudara u pot-

punosti sa sekvencom iz PDB, one su poravnate koristeći alat Kalign [37]. Ostatku mreže prosljeđuju se četiri predloška za koja je izračunat najveći očekivani broj ispravno poravnatih rezidua (izlazna varijabla „sum_probs" alata HHSearch).



Slika 3.4 Strukture mioglobina iz četiri različita organizma, preuzeto iz [13]

Misao vodilja iza strukturne pretrage može se pronaći u staroj izreci: „nihil sub sōle novum" [38] - nema ništa novo pod Suncem. Ideja jest da, iako proteini mutiraju i evoluiraju, strukture srodnih proteina ostaju relativno slične [10]. Primjer toga može se vidjeti na slici 3.4 koja sadrži tercijarne strukture mioglobina iz četiri različita organizma, redom u smjeru kazaljke na satu počevši od gornje desne: čovjeka, afričkog slona, crnoperajne tune i goluba. Vidljivo je da su sve četiri varijante strukturno vrlo slične, iako nemaju iste lance aminokiselina: sekvenca mioglobina afričkog slona ima 80% podudaranja s mioglobinom ljudskog, mioglobin tune 45% a goluba samo 25% podudaranja s ljudskim. Usprkos različitim lancima, funkcija proteina ekvivalentna je u sva četiri organizma pa stoga ni strukture nisu drastično izmijenjene [13].

Poglavlje 4

Evoformer

Centralni dio AlphaFold 2 mreže jest Evoformer - transformator kojemu je zadatak maksimizacija količine korisnih informacija koje se mogu izvući iz MSA i strukturnih predložaka [13]. Evoformer predviđanje strukture proteina razmatra kao problem zaključivanja grafa (*graph inference*) u trodimenzionalnom prostoru, gdje su bridovi grafa definirani bliskim reziduama u lancu [8].

4.1 Pretprocesiranje

Prije nego se podaci procesiraju u samoj mreži, izvodi se nekoliko koraka pretprocesiranja kako bi se optimizirale performanse mreže.

MSA grupiranje

S obzirom na to da najgora vremenska i memorijska složenost Evoformer mreže iznosi $N_{seq}^2 * N_{res}$, gdje N_{seq} predstavlja broj sekvenci (redaka), a N_{res} broj rezidua (stupaca) u MSA, provodi se grupiranje MSA (eng. *MSA clustering*) kako bi se smanjile dimenzije MSA korištenog tijekom izračuna uz što manji negativni učinak na točnost mreže [8]. Grupiranje podrazumijeva biranje nasumičnog podskupa fiksne veličine N_{clust} sekvenci koje predstavljaju reprezentativan skup, i svaku se sekvencu iz MSA asocira s njoj najbližom sekvencom iz reprezentativnog skupa. Kako bi se

zadržala ograničena veličina, uzimaju se u obzir samo frekvencije aminokiselina i brisanja (iz PHMM stanja M i D iz Poglavlja 3.1) svih sekvenci asociranih s reprezentativnim sekvencama, te se one koriste kao dodatna obilježja (*extra features*) uz reprezentativne sekvence. Time se postiže gotovo dvostruki broj obilježja po sekvenci bez udvostručenja broja sekvenci [8].

Procedura MSA grupiranja jest sljedeća:

1. Nasumično se odabere N_{clust} sekvenci koje služe kao središta grupa (*clustera*), pri čemu sekvenca čija se konačna struktura traži predstavlja središte prve grupe.
2. Generira se maska takva da svaka pozicija u centru MSA grupe ima vjerojatnost od 15% da je uključena u toj maski. Svaki element MSA koji jest uključen u maski zamijeni se uz sljedeće uvjete:
   - u 10% slučajeva zamijeni se uniformno uzorkovanom aminokiselinom
   - u 10% slučajeva zamijeni se aminokiselinom uzorkovanom iz profila MSA na danoj poziciji
   - u 10% slučajeva se ne zamijeni
   - u 70% slučajeva zamijeni se posebnim znakom (*masked_msa_token*).
3. Preostale se sekvence pridruže grupa koje su im najbliže po Hammingovoj udaljenosti [39] ignorirajući zamaskirane rezidue i praznine. Za svaku se grupu računa nekoliko statističkih vrijednosti (npr. razdioba aminokiselina po svakom stupcu).
4. Od sekvenci iz MSA koje nisu izabrane kao središta grupa u prvom koraku nasumično se odabere N_{extra_seq} sekvenci bez zamjenijenjenih aminokiselina (u slučaju da takvih sekvenci ima manje od N_{extra_seq}, odaberu se sve) koje čine dodatna MSA obilježja (*extra_msa_feat* u Tablici 4.1).

Određivanje obilježja

Glavni ulazi u model navedeni su u Tablici 4.1, a sva obilježja korištena za izračun glavnih ulaza, kao i njihove dimenzije, navedena su u Dodatku A.

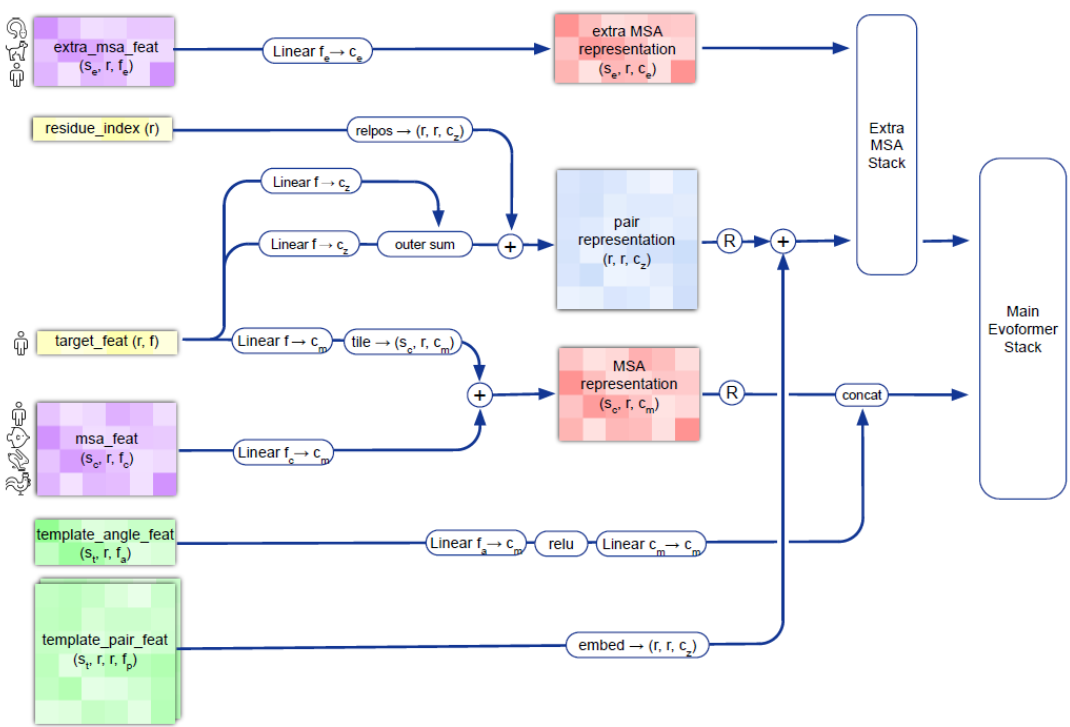
Tablica 4.1 Glavni ulazi Evoformer modula

Obilježje Oblik	Opis
target_feat [N_{res}, 21]	Sadrži „aatype”
residue_index [N_{res}, 21]	Sadrži „residue_index”
msa_feat [N_{clust}, N_{res}, 49]	Stvara se konkatenacijom „cluster_msa”, „cluster_has_deletion”, „cluster_deletion_value”, „cluster_deletion_mean” i „cluster_profile”. Uzme se $N_{cycle} * N_{ensemble}$ nasumičnih uzoraka iz ovog obilježja kako bi se svakoj iteraciji mreže pružio različit uzorak
extra_msa_feat [N_{extra_seq}, N_{res}, 25]	Stvara se konkatenacijom „extra_msa”, „extra_msa_has_deletion” i „extra_msa_deletion_value”. Kao i „msa_feat”, nasumično se uzorkuje $N_{cycle} * N_{ensemble}$ puta
template_pair_feat [N_{templ}, N_{res}, N_{res}, 88]	Stvara se konkatenacijom „template_distogram” i „template_unit_vector” te nekoliko obilježja rezidua, koje se pretvore u obilježja parova. Također, pomoću postupaka *tiling* i *stacking* u oba smjera rezidua uključen je „template_aatype”. Uključene su i maske „template_pseudo_beta_mask” i „template_backbone_frame_mask”, pritom je obilježje $f_{ij} = mask_i * mask_j$
template_angle_feat [N_{templ}, N_{res}, 51]	Stvara se konkatenacijom „template_aatype”, „template_torsion_angles”, „template_alt_torsion_angles” i „template_torsion_angles_mask”

Dimenzije korištene u Tablici 4.1:

- N_{res} - broj rezidua
- N_{clust} - broj grupa MSA
- N_{extra_seq} - broj sekvenci koje nisu ni u jednoj grupi MSA
- N_{templ} - broj strukturnih predložaka
- N_{cycle} - broj iteracija cijele mreže
- $N_{ensemble}$ - broj iteracija Evoformera

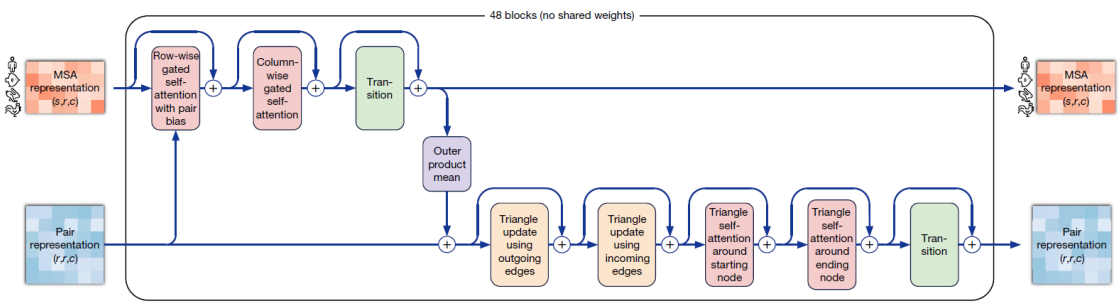
Ugrađivanje ulaznih obilježja za obradu u Evoformeru prikazano je na Slici 4.1.



Slika 4.1 Dijagram ugrađivanja ulaznih obilježja, preuzeto iz [8]

4.2 Mreža

Evoformer mreža ima arhitekturu dvotoranjskog (*two-tower*) transformatora s osnom samopozornošću (*axial self-attention*) u MSA stogu te trokutnim multiplikativnim ažuriranjem (*triangular multiplicative update*) i trokutnom samopozornosšću (*triangular self-attention*) u stogu parova. Kako bi se omogućila komunikacija između stogova, koristi se srednji vektorski umnožak i podešavanje pozornosti (*attention biasing*) [8]. Mreža se sastoji od $N_{block} = 48$ blokova koji ne dijele težine (*attention weight*), od kojih svaki uzima MSA reprezentaciju i reprezentaciju parova kao ulaz te procesira u nekoliko slojeva. Izlaz svakog sloja dodaje se trenutnim reprezentacijama kroz rezidualnu vezu. Pritom izlazi nekih slojeva prije dodavanja prođu kroz Dropout [40], kako bi se smanjio overfitting. Posljednji blok Evoformera daje obrađene i rafinirane reprezentacije MSA i parovima koje sadrže informacije potrebne strukturnom modulu i pomoćnim glavama mreže. Uz to se predaje i reprezentacija jednom sekvencom koja nastaje linearnom projekcijom prvog retka reprezentacije MSA [8]. Prikaz bloka Evoformera nalazi se na Slici 4.2. Dvotoranjska arhitektura podrazumijeva zasebnu obradu reprezentacije MSA u MSA stogu i reprezentacije parova u stogu parova uz izmjenjivanje i kombiniranje informacija između stogova.



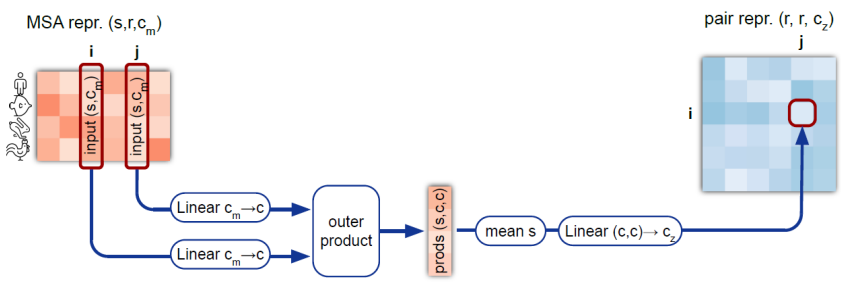
Slika 4.2 Dijagram bloka Evoformera, preuzeto iz [8]

4.2.1 MSA stog

Osna samopozornost MSA stoga (*MSA stack*) odnosi se na računanja težina pozornosti posebno po retcima i stupcima. Samopozornost po retcima (*MSA row-wise gated self-attention with pair bias*) gradi težine pozornosti za parove rezidua i pritom koristi informacije iz reprezentacije parovima. Tim „podešavanjem" reprezentacije MSA podacima iz reprezentacije parovima potiče se konzistentnost između reprezentacije MSA i reprezentacije parova. Samopozornost po stupcima (*MSA column-wise gated self-attention*) omogućava razmjenu informacija između elemenata koji pripadaju istoj rezidui (istom stupcu). Izračunima samopozornosti učetverostruči se broj kanala u MSA reprezentaciji, pa se kroz *MSA transition* broj kanala reducira na početni.

4.2.2 Srednji vektorski umnožak

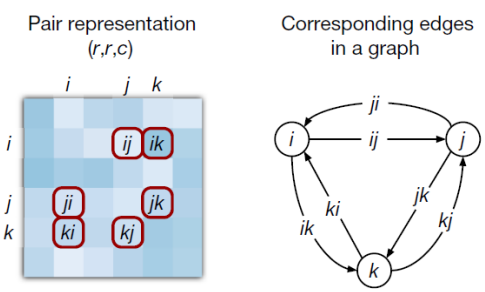
Nakon obrade u MSA stogu, podaci iz reprezentacije MSA koriste se za ažuriranje reprezentacije parovima kroz srednji vektorski umnožak (*outer product mean*) prikazan na Slici 4.3. Sve sekvence iz MSA linearno se projiciraju kako bi se smanjio broj kanala, te se za svaki par rezidua na pozicijama (i, j) računa vektorski produkt stupaca (i, j) i usrednji po sekvencama. Ta se usrednjena vrijednost koristi za ažuriranje reprezentacije parovima na mjestu (i, j). Ovaj je korak memorijski zahtjevan jer je za izračun potrebno stvarati tenzore s puno dimenzija.



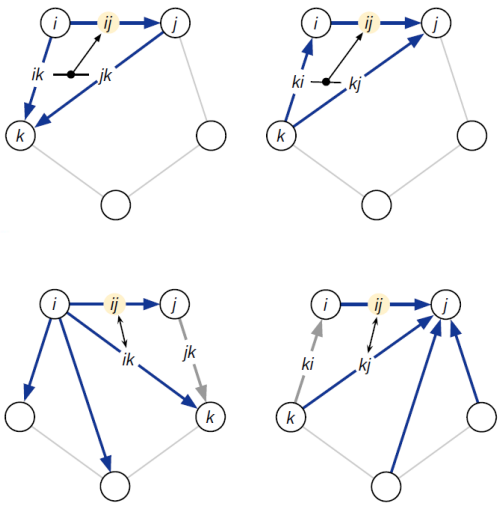
Slika 4.3 Proces ažuriranja reprezentacije parovima reprezentacijom MSA, preuzeto iz [8]

4.2.3 Stog parova

Nakon što je reprezentacija parovima ažurirana od strane MSA stoga, kreće obrada u stogu parova (*pair stack*). Ovdje se reprezentacija parovima interpretira kao usmjereni graf, kao što je prikazano na Slici 4.4a. Pritom je ključno što se teži k zadovoljavanju nejednakosti trokuta među trojkama rezidua. Prvi korak obrade u stogu parova jest trokutno multiplikativno ažuriranje. *Triangular multiplicative update* ažurira reprezentaciju parovima tako da kombinira informacije iz svakog trokuta sačinjenog od bridova grafa ij, ik i jk. Svaki brid ij ažurira se informacijama iz druga dva brida (ik i jk), a pritom se računaju dvije simetrične verzije, za *outgoing* i *incoming* bridove (*outgoing* je npr. ik a *incoming* ki). Gornja dva koraka na Slici 4.4b prikazuju proces trokutnog multiplikativnog ažuriranja, pritom prvi dijagram prikazuje varijantu za *outgoing* bridove, a drugi za *incoming* bridove.



(a) Lijevo: reprezentacija parovima
Desno: graf reprezentacije parovima



(b) Triangular multiplicative update i Triangular self attention

Slika 4.4 Dijelovi stoga parova, preuzeto iz [8]

Drugi korak obrade u stogu parova jest trokutna samopozornost. *Triangular self-attention* ažurira reprezentaciju parovima tako da početni čvor i ažurira brid ij vrijednostima svih bridova koji imaju isti početni čvor (svi bridovi ik gdje $k \neq i, j$). Na to hoće li neki brid ik ažurirati brid ij osim same sličnosti bridova utječu i informacije dobivene iz trećeg brida jk. Računa se i simetričan proces za krajnje čvorove (dakle ki i kj umjesto ik i jk). Slično kao i u MSA stogu, učetverostruči se broj kanala u reprezentaciji parovima, pa se kroz *pair transition* broj kanal reducira na početni.

4.3 Dodatni ulazi

Pored glavnog dijela mreže kojeg čine MSA stog i stog parova, za postizanje veće točnosti koriste se i stog predložaka (*template stack*) i dodatan MSA stog (*unclustered MSA stack*) [8].

4.3.1 Stog predložaka

Obilježja iz strukturnih predložaka linearno se projiciraju kako bi se stvorile reprezentacije predlošcima. Svaka reprezentacija predlošcima nezavisno se procesira u stogu predložaka i izlazne se reprezentacije agregiraju pomoću *template point-wise attention*, gdje se na temelju reprezentacija parova grade upiti i nadziru individualni predlošci. Izlazi stoga predložaka dodaju se reprezentaciji parovima. Kutovi torzije ugrađuju se pomoću malog višeslojnog perceptrona (Multi-layer Perceptron (MLP)) i konkateniraju s MSA reprezentacijama kao dodatni redovi sekvenci s drugačijim skupom težina u odnosu na sekvence iz MSA. Dodatni redovi sudjeluju u svim operacijama MSA stoga.

4.3.2 Dodatni MSA stog

U ovom se stogu sekvence koje ne pripadaju nijednoj grupi MSA ugrađuju u posebne reprezentacije MSA koje se procesiraju u dodatnom MSA stogu (*Extra MSA stack*) koji se sastoji od četiri bloka slična onima iz Evoformera. Ključna razlika

extra MSA stoga u odnosu na Evoformer je što se u *extra MSA* stogu koristi globalna samopozornost po stupcima te što su same reprezentacije manje veličine, što omogućava brže procesiranje većeg broja sekvenci. Konačne reprezentacije parovima koriste se kao dodatni ulazi u glavni Evoformer stog, dok se reprezentacije MSA ne koriste u daljnjim izračunima.

Poglavlje 5

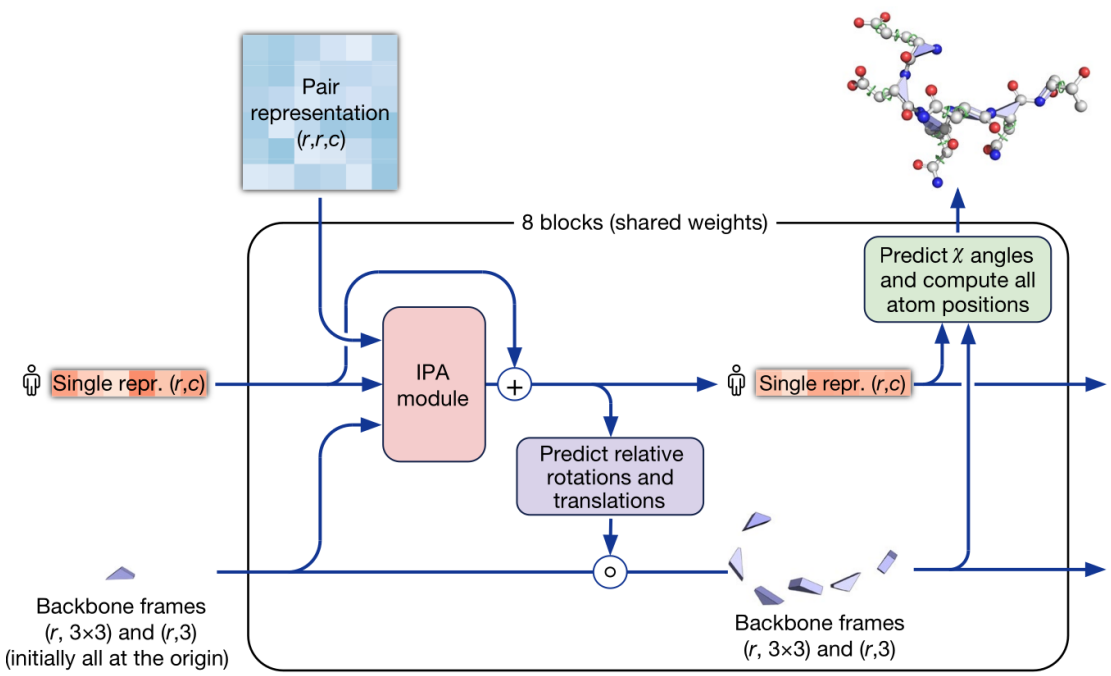
Strukturni modul

Posljednji dio AlphaFold 2 mreže jest strkutkruni modul (*Structure module*). Ovaj dio mreže uzima apstraktne reprezentacije proteina koje dobije iz Evoformera te iz njih gradi konkretnu tercijarnu strukturu proteina, odnosno pretvara procesiranu MSA reprezentaciju i reprezentaciju parovima u trodimenzionalne koordinate svih atoma u lancu proteina [8, 13].

5.1 Mreža

Strukturni modul sastoji se od osam blokova koji međusobno dijele težine. Za glavni ulaz uzima reprezentaciju jednom sekvencom iz Evoformera (*single representation*). Reprezentacijom parovima podešavaju se mape afiniteta za određivanje pozornosti. U svakom se bloku strukturnog modula, prikazanom na Slici 5.1, ažuriraju apstraktna reprezentacija jednom sekvencom kao i trodimenzionalna reprezentacija, koja je kodirana kao jedan okvir okosnice (*backbone frame*) po rezidui. Aminokiseline u lancu modeliraju se trokutima koje čine tri atoma okosnice aminokiseline, i ti trokuti (okviri) slobodno *plutaju* prostorom, pa trodimenzionalna reprezentacija počinje kao „oblak rezidua" (*residue gas*). Obradom u strukturnom modulu ti se okviri (trokuti) slažu u prostoru kako bi se formirala tercijarna struktura. Okviri su predstavljeni euklidskim transformacijama koje su određene uređenim parovima rotacije i translacije $T_i := (R_i, \vec{\boldsymbol{t}}_i)$ gdje je R_i rotacija u trodimenzionalnom pros-

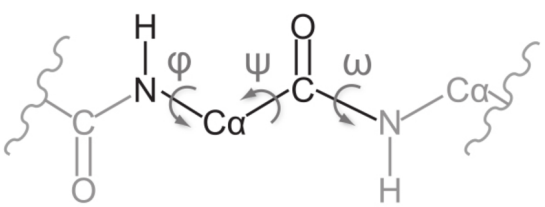
toru a $\vec{\boldsymbol{t}}_i$ vektor translacije u trodimenzionalnom prostoru. Takav uređeni par služi kao euklidska transformacija (transformacija krutog tijela) iz lokalnog referentnog sustava (okvir okosnice) u globalni referentni sustav (struktura cijelog proteina), odnosno transformira lokalne koordinate $\vec{x}_l$ u globalne koordinate $\vec{x}_g$ Hadamardovim produktom $\vec{x}_g = T_i \circ \vec{x}_l = R_i\vec{x}_l + \vec{\boldsymbol{t}}_i$. Pritom se euklidska rotacija u trodimenzionalnom prostoru ne vrši 3×3 matricom transformacije kako je uobičajeno, već kvaternionima, koji koriste samo četiri komponente u odnosu na devet nužnih za matricu rotacije pa su stoga računski i memorijski manje zahtjevni te su ujedno i prikladniji za manje kutove zbog manjeg rizika od gubitka preciznosti (*underflowa*) [41, 42]. U početku se okviri okosnice inicijaliziraju transformacijom identiteta kojom se smjeste u ishodište globalnog koordinatnog sustava. Autori AlphaFold 2 nazvali su to „inicijalizacija crnom rupom" (*black hole initialization*). Time su svi okviri okosnice smješteni u istoj točki u trodimenzionalnom prostoru i imaju istu orijentaciju.



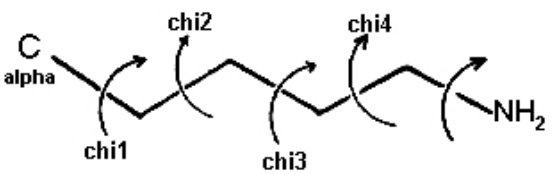
Slika 5.1 Dijagram bloka strukturnog modula, preuzeto iz [8]

U svakom bloku strukturnog modula procesiranje započinje ažuriranjem reprezentacije jednom sekvencom u Invariant Point Attention (IPA) modulu koji je zaslužan za određivanje pozornosti u kontekstu skupa okvira (euklidskih transformacija) na način da ta pozornost nije pod utjecajem globalnih euklidskih transformacija nad tim okvirima. IPA modul se brine za to da su ažuriranja u globalnom sustavu ekvivalentna onima u lokalnom i obrnuto, odnosno da ista euklidska transformacija jednako utječe na okvire i u lokalnom i u globalnom kontekstu. Nakon toga se reprezentacija jednom sekvencom mapira na okvire za ažuriranje okosnice. Ti su okviri isto euklidske transformacije u trodimenzionalnom prostoru sačinjene od rotacije i translacije, te se njima okviri okosnice kontinuirano nadograđuju.

Kako bi se odredile koordinate svih atoma, svaka se rezidua parametrizira po kutovima torzije $\{\omega,\ \phi,\ \psi,\ \chi_1,\ ...,\ \chi_4\}$, koji u tom slučaju bivaju jedini stupnjevi slobode, jer su kutovi i udaljenosti između atoma aminokiseline fiksni. Pritom se atomi grupiraju u čvrste grupe (*rigid groups*) ovisno o njihovoj zavisnosti o kutovima torzije. Definirane su čvrste grupe za tri kuta torzije okosnice $\{\omega,\ \phi,\ \psi\}$ (Slika 5.2a) i četiri kuta torzije bočnog lanca $\{\chi_1,\ \chi_2,\ \chi_3,\ \chi_4\}$ (Slika 5.2b).



(a) Kutovi torzije okosnice, preuzeto iz [43]



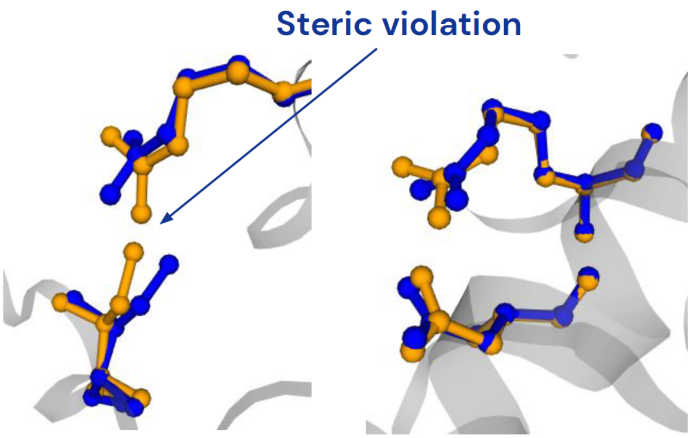
(b) Kutovi torzije bočnog lanca, preuzeto iz [44]

Slika 5.2 Kutovi torzije aminokiseline

Za predviđanje kutova torzije, zaslužna je plitka rezidualna neuronska mreža (*ResNet*) [8, 45]. Kutovi torzije prikazani su vektorima koji sadrže dvodimenzionalne koordinate na jediničnoj kružnici kako bi se olakšao izračun i stvaranje matrica rotacije te eliminiralo potencijalno izlaženje iz intervala $[0, 2\pi]$ radijana. Nakon što se kutovi torzije izračunaju, pretvore se u okvire za čvrste grupe atoma i primijene na pripadne strukture aminokiselina.

5.2 Relaksacija

Konačni rezultat svih slojeva strukturnog modula ne mora nužno zadovoljavati sva stereokemijska ograničenja, pa je zato potrebno provesti relaksaciju strukture proteina. U AlphaFold 2 relaksacija se provodi nakon izvođenja cijelog strukturnog modula koristeći alat OpenMM [46] na AMBER99SB polje sila [47]. Sama relaksacija provodi se iterativnom ograničenom procedurom minimizacije energije (eng. *iterative restrained energy minimization procedure*). U svakoj se relaksaciji minimizira AMBER99SB polje sila koje je dodatno harmonički ograničeno kako sustav ne bi previše odstupao od početne strukture. Kada minimizator konvergira, odredi se koje rezidue još uvijek ne zadovoljavaju stereokemijska ograničenja, pa se iz tih rezidua maknu ograničenja sa svih atoma i ponovno provede minimizacija, ovaj put koristeći već minimiziranu strukturu iz prethodne iteracije. Taj se proces ponavlja sve dok se ne razriješe sva kršenja stereokemijskih ograničenja. Na kraju se potpuna minimizacija energije i smještanje atoma vodika provodi OpenMM simulacijskim paketom, koristeći predefinirane vrijednosti alata [8].



Slika 5.3 Prikaz relaksacije strukture proteina, preuzeto iz [48]

Primjer relaksacije vidljiv je na Slici 5.3, gdje je narančastom bojom prikazan struktura prije relaksacije, na kojoj je strelicom označeno kršenje stereokemijskih ograničenja. To se kršenje razriješi relaksacijom, čiji je konačan rezultat prikazan plavom bojom.

5.3 Recikliranje

Konačni rezultati strukturnog modula nakon relaksacije, kao i dodatni izlazi i funkcije gubitka, šalju se tri puta na ulaz Evoformera tako da se cijela mreža izvršava ukupno četiri puta. Time se postiže produbljivanje mreže i povećava preciznost procesiranjem više verzija ulaznih obilježja bez da se značajno poveća vrijeme treniranja ili broj parametara. Tijekom zaključivanja, recikliranje pretvara AlphaFold 2 u ponavljajuću neuronsku mrežu (Recurrent Neural Network (RNN)) s dijeljenim težinama, gdje svaka iteracija kao ulaze uzima izlaze iz prethodne (za prvu iteraciju oni su nepostojeći, što mreži ne predstavlja problem), ali i generira nove ulaze kroz nasumično uzorkovanje MSA i grupiranje MSA [8].

Poglavlje 6

Pouzdanost

6.1 Metrike

U svrhu određivanja točnosti i preciznosti modela za predviđanje strukture proteina, koriste se određene metrike sličnosti struktura proteina. Najčešće korištene su standardna devijacija pozicija atoma (RMSD), Global Distance Test (GDT), i Local Distance Difference Test (lDDT).

Najjednostavniji i najdirektniji pokazatelj sličnosti dvije strukture proteina jest standardna devijacija pozicija atoma (*root-mean-square deviation of atomic positions*). Računa se kao srednja vrijednost svih udaljenosti između ekvivalentnih atoma strukturno poravnatih proteina, izrazom 6.1, gdje je N broj atoma u strukturi a $\delta_i = |u_i - v_i|^2$ udaljenost između i-tih atoma u i v na poravnatim strukturama, najčešće izražena u angstremima (Å) [49].

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{6.1}$$

RMSD daje prosječno odstupanje između struktura poravnatih proteina, te se najčešće koristi RMSD atoma okosnice, konkretno Cα atoma svake rezidue. Među glav-

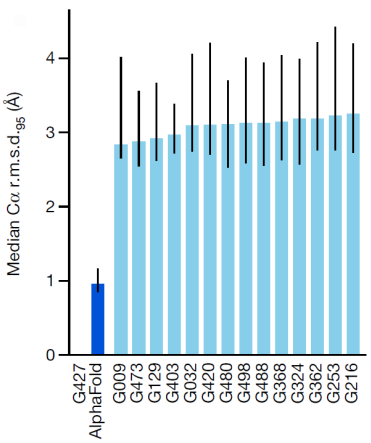
nim problemima RMSD jest mogućnost da regije sa značajnijim odstupanjima imaju prevelik utjecaj na konačnu vrijednost u odnosu na regije koje se više podudaraju, pa ne daje uvijek ispravnu sliku o strukturnoj sličnosti proteina. Jedna od češćih metoda poravnanja proteina jest upravo translacija i rotacija struktura s ciljem minimizirnja RMSD-a.

GDT zamišljen je kao evolucija RMSD metrike, originalno implementiran u Local-Global Aligment (LGA) alatu [50]. Pomoću GDT želi se ublažiti osjetljivost RMSD na regije koje značajno odstupaju između dvije poravnate strukture. GDT nema mjernu jedinicu, već se izražava postotkom (od nula do sto) i računa nad Cα atomima okosnice. Sličnost iterativno poravnatih struktura „boduje" se u dvadeset *cutoff* kategorija udaljenosti, od 0,5 Å do 10 Å. Strukture su sličnije što je GDT veći, odnosno što je više „bodova" unutar što ranijeg *cutoffa*. Na CASP natjecanju koristi se modificirana verzija zvana GDT_TS (Global Distance Test Total Score) koja ima četiri *cutoff* udaljenosti: 1, 2, 4 i 8 Å [51].

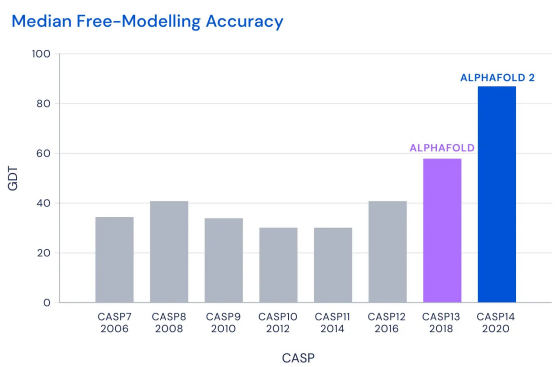
Za razliku od RMSD i GDT koje zahtijevaju da strukture budu poravnate i primarno promatraju samo Cα atome okosnice, lDDT je osmišljen da boduje sličnost struktura bez potrebe za globalnim poravnanjem struktura te promatra sve atome na strukturama, ne samo atome okosnice. Važan aspekt lDDT jest i mogućnost evaluacije proteina koji se sastoje od više domena bez potrebe za prethodnom obradom [52]. LDDT primarno mjeri koliko su vjerno lokalne interakcije očuvane u strukturi čija se sličnost promatra u odnosu na referentnu strukturu. Računa se za sve parove atoma u referentnoj strukturi čije su udaljenosti manje od predefinirane granice (radijusa inkluzije), a koji ne pripadaju istoj rezidui. Pritom se računa za više vrijednosti radijusa inkluzije, tako da se promatra koliki je udio udaljenosti ostao očuvan na strukturi koja se evaluira u odnosu na referentnu strukturu. Konačan rezultat računa se kao prosječni udio očuvanih udaljenosti unutar četiri radijusa inkluzije: 4, 2, 1 i 0,5 Å, i iskazan je kao postotak (poprima vrijednosti između nula i sto) [52].

6.2 AlphaFold 2 i CASP14

Na četrnaestom CASP natjecanju održanom 2020. godine AlphaFold 2 model pokazao se značajno točnijim od modela ostalih natjecatelja. Medijan odstupanja atoma okosnice (Cα) iznosi 0,96 Å za interval pouzdanosti 95% između 0,85 i 1,16 Å, a medijan odstupanja svih atoma iznosi 1,5 Å za interval pouzdanosti 95%[1] između 1,2 i 1,6 Å, dok idući najbolji model ima medijan odstupanja atoma okosnice od 2,8 Å za interval pouzdanosti 95% između 2,7 i 4,0 Å i medijan odstupanja svih atoma od 3,5 Å za interval pouzdanosti 95% između 3,1 i 4,2 Å, kao što je prikazano na Slici 6.1a [8]. Na svim domenama u CASP14 natjecanju, prosječni GDT_TS za AlphaFold 2 iznosio je 87,32 za top pet predanih modela po domeni i 88,01 za najbolji predani model po domeni, dok je medijan GDT_TS za sve domene iznosio 92,4. [9, 48].



(a) RMSD atoma okosnice najboljih petnaest modela s CASP14, preuzeto iz [8]



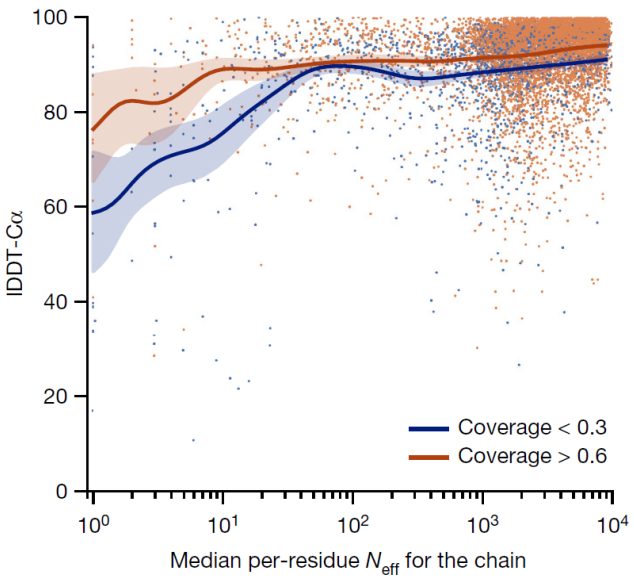
(b) Točnost pobjedničkih modela na CASP natjecanju između 2006. i 2020., preuzeto iz [54]

Slika 6.1 Performanse AlphaFold 2 na CASP natjecanju

AlphaFold 2 pokazao se podjednako točnim i za kraće i za dulje sekvence, i postigao značajnu preciznost i za novootkrivene strukture, netom dodane u PDB bazu [8,

[1]Interval pouzdanosti od 95% znači da 95% svih vrijednosti nekog skupa pada unutar tog intervala [53]

9]. Autori AlphaFold 2 uočili su da veličina MSA ima značajan utjecaj na točnost predviđenog modela, konkretno da su rezultati predviđanja kod kojih MSA sadrži manje od tridesetak sekvenci značajno manje točni [8, 9, 12]. Odnos veličine MSA i pouzdanosti predviđanja prikazan je na Slici 6.2, gdje je na horizontalnoj osi smješten red veličine MSA, a na vertikalnoj osi lDDT za atome okosnice. Pad točnosti značajniji je u slučajevima gdje strukturni predlošci pokrivaju mali dio početne sekvence, tako su ovdje rezultati kod kojih predlošci pokrivaju manje od 30% ulazne sekvence prikazani plavom bojom, a narančastom rezultati kod kojih predlošci pokrivaju više od 60% ulazne sekvence [8]. Utjecaj manjka pokrivenosti predlošcima opada kada MSA ima barem stotinjak sekvenci.



Slika 6.2 Odnos veličine MSA, pokrivenosti predlošcima i lDDT atoma okosnice, preuzeto iz [8]

Jedno od značajnijih ograničenja AlphaFold 2 modela jest lošije predviđanje intrinzično neuređenih regija proteina, koje sačinjavaju preko 30% sekvenci ljudskog proteoma duljih od trideset aminokiselina [55, 56]. Određivanje strukture intrinzično neuređenih proteina općenito je iznimno zahtjevno, i generalno se pridaje veći značaj jasno definiranim strukturama, za koje se AlphaFold 2 pokazao iznimno sposoban [11, 12, 57].

Poglavlje 7

Rezultati

7.1 Okruženje

Za pokretanje predviđanja AlphaFold 2 modelom korišten je računalni klaster Isabella Sveučilišnog računskog centra Sveučilišta u Zagrebu. Klaster Isabella nastao je u siječnju 2002. godine, te je u svibnju iste godine stavljen na raspolaganje akademskoj zajednici [58]. Danas se klaster sastoji od 135 računalnih čvorova s preko 260 CPU procesora, 3.100 CPU procesorskih jezgri, 12 grafičkih procesora i 16 TB radne memorije. Na svim čvorovima nalaze se po dva Intel Xeon CPU procesora, a čvorovi s grafičkim procesorima imaju po tri NVIDIA Tesla V100 grafička procesora [59]. Na klasteru je, između ostalog, unaprijed instaliran i postavljen AlphaFold 2 program, tako da korisnik ne treba vršiti nikakvu konfiguraciju. U kontekstu ovog rada, koristila se 2.1.1 non-docker verzija AlphaFold 2 [60]. Klasteru se iz CARNET mreže pristupa preko pristupnog poslužitelja teran.srce.hr koristeći Secure Shell (SSH) protokol [59].

7.2 Strukture

U svrhu evaluacije primjenjivosti AlphaFold 2 modela za predviđanje peptida, korištene su kraće sekvence (između 36 i 223 aminokiseline) preuzete iz RCSB PDB baze proteina [61]. PDB oznake sekvenci korištenih za evaluaciju su: 1BBA [62],

1UCS [63], 1V3Y [64], 2OS3 [65], 3NIR [66] i 4I8H [67]. U Tablici 7.1 prikazani su podaci o predviđanjima AlphaFold 2 modela za svaku od prethodno navedenih sekvenci, izračunate metrike točnosti objašnjene u Poglavlju 6.1 i broj aminokiselina u lancu svake sekvence. Za izračun GDT korišten je LGA alat [50] dostupan na proteinmodel.org, za izračun lDDT korišten je lDDT alat [52] dostupan swissmodel.expasy.org, a za izračun RMSD korišten je ChimeraX alat razvijen na Sveučilištu u Kaliforniji, San Francisco [68].

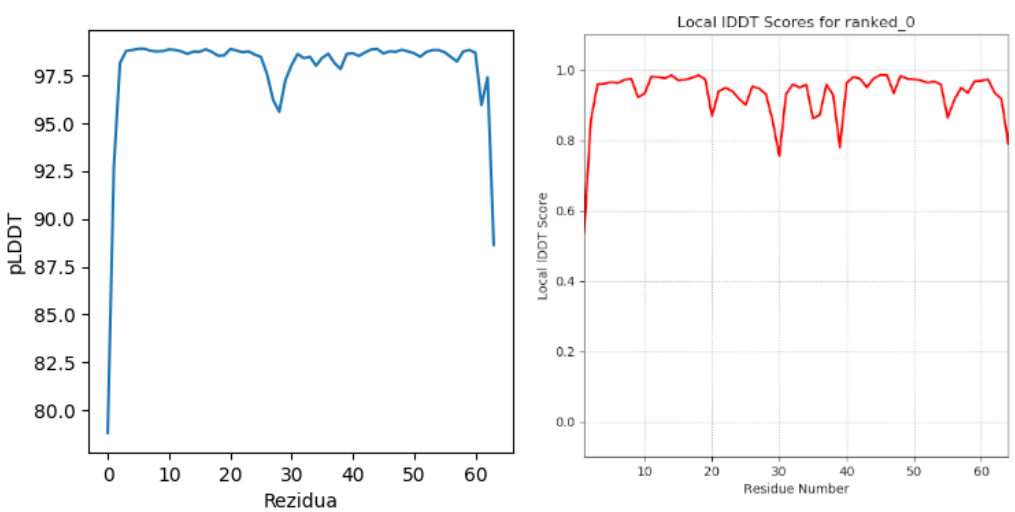
Tablica 7.1 Rezultati predviđanja AlphaFold 2 modelom

Peptid	Duljina lanca	Veličina MSA	Broj predložaka	RMSD	GDT	lDDT	Vrijeme izvršavanja
1BBA	36	220	20	2,06 Å	85,42	70,80	35 min
3NIR	46	221	7	0,96 Å	98,91	93,10	40 min
1UCS	64	1206	20	0,47 Å	99,61	93,72	37 min
2PNE	81	1719	7	42,83 Å	19,21	17,70	45 min
1V3Y	192	12333	20	0,52 Å	95,60	91,90	46 min
2OS3	205	12030	20	0,72 Å	97,32	94,10	48 min
4I8H	223	11657	20	0,46 Å	98,99	95,99	53 min

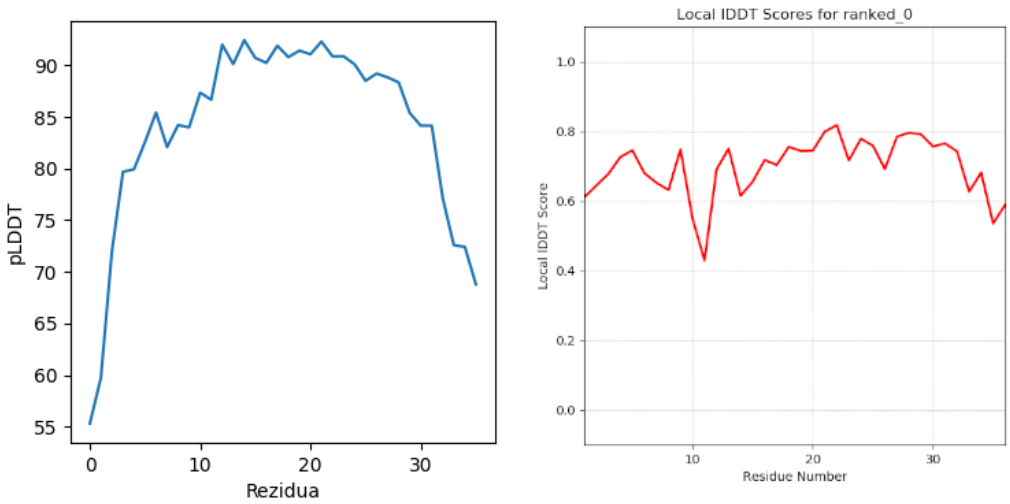
Na vrijeme izvršavanja najveći utjecaj ima faza pretraživanja i pretprocesiranja ulaza, zatim relaksacija i tek na kraju samo zaključivanje strukture. Prema autorima AlphaFold 2 modela, vrijeme samog zaključivanja na jednom V100 grafičkom procesoru sa 16 GB memorije iznosi oko 36 sekundi za lanac od 256 aminokiselina, 66 sekundi za lanac s 384 aminokiseline te više od dva sata za lance s oko 2.500 aminokiselina. Veći problem predstavlja memorijska složenost, gdje se više od 16 GB memorije može prijeći kod lanaca s više od 1.300 aminokiselina [8].

Među predviđanjima korištenih sekvenci, pet od sedam struktura postiglo je visoku točnost - RMSD manji od 1 Å te GDT i lDDT veći od 90. Nešto manje točna struktura predviđena je za najkraću sekvencu, 1BBA, a za 2PNE je struktura gotovo u potpunosti promašena. Najveći razlozi za nedostatak točnosti u slučaju te dvije sekvence jesu niža podudaranja sekvenci iz MSA i ulaznih sekvenci, te nedostatak kvalitetnih i reprezentativnih strukturnih predložaka. Ukratko, u oba je slučaja pro-

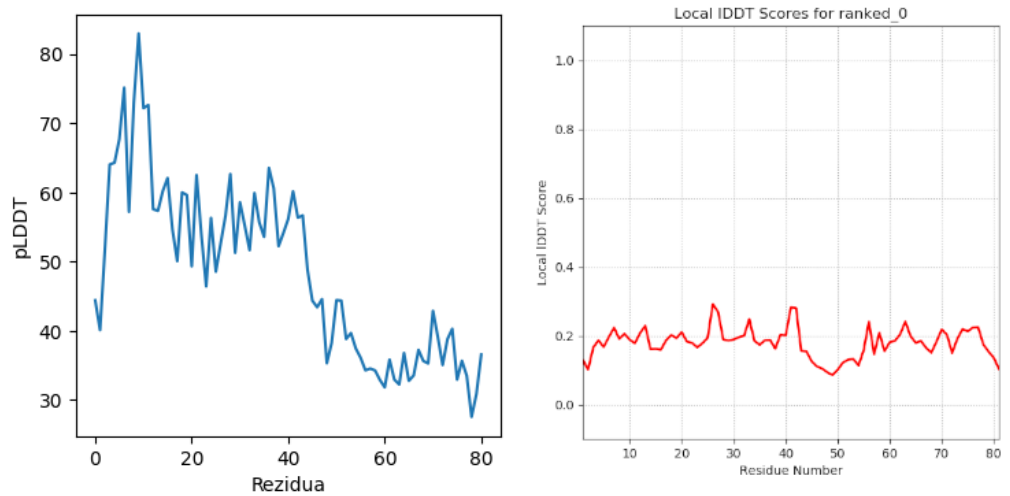
blem što nije poznato dovoljno srodnih struktura, pa AlphaFold 2 nije imao dovoljno „predznanja" za kvalitetno predviđanje. Pritom je korisno što AlphaFold 2 na temelju (ne)dostupnih podataka prognozira gdje će predviđanja imati manju točnost, i te se prognoze uvelike podudaraju s odstupanjima izračunata alatima za evaluaciju predviđene strukture. Za svaku predviđenu strukturu AlphaFold 2 generira očekivano odstupanje strukture za svaku reziduu u lancu, predstavljeno plDDT (predicted lDDT) vrijednostima. U većini slučajeva vrijednost plDDT nije u potpunosti mjerodavna s obzirom na točnost konačne strukture i izračunati lDDT, ali generalno ispravno odredi koje će regije biti manje precizne. Globalna vrijednost plDDT za neku strukturu većinom bude optimistična u odnosu na točnost predviđene strukture, dok vrijednosti za regije koje prognozira manju točnost bivaju pesimističnije u odnosu na izračunati lDDT. Autori AlphaFold 2 odredili su kategorije pouzdanosti modela temeljene na vrijednostima plDDT-a: jako niska za manje od 50 niska između 50 i 70, visoka između 70 i 90 te vrlo visoka iznad 90 [8]. Prema tome, sve regije za koje je plDDT veći od 70 trebale bi biti ispravno predviđene. Na Slikama 7.1, 7.2, 7.3 prikazani su grafovi lDDT za sekvence 1UCS, 1BBA i 2PNE, gdje lijevi grafovi predstavljaju plDDT kojeg je prognozirao AlphaFold 2, a desni je izračunat u odnosu na referentnu strukturu iz PDB baze koristeći lDDT alat [52]. Na Slici 7.4 prikazana su poravnanja predviđenih i stvarnih struktura sekvenci 1UCS, 1BBA i 2PNE izrađene u alatu ChimeraX [68].



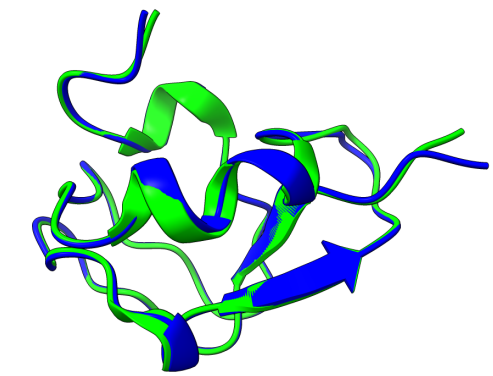
Slika 7.1 Usporedba plDDT i izračunatog lDDT za protein 1UCS



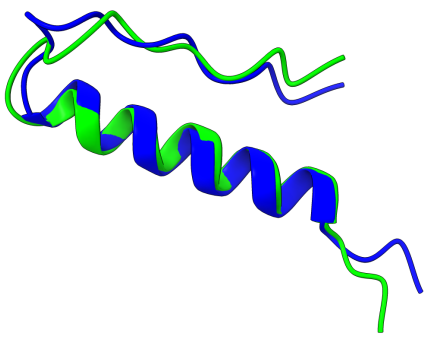
Slika 7.2 Usporedba plDDT i izračunatog lDDT za protein 1BBA



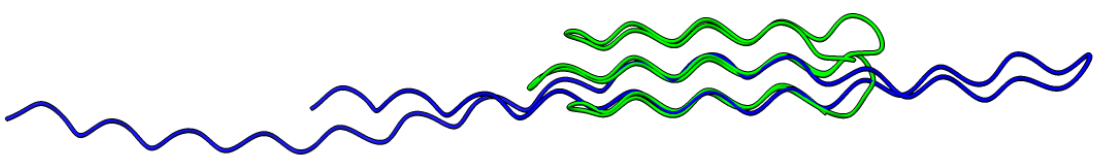
Slika 7.3 Usporedba plDDT i izračunatog lDDT za protein 2PNE



(a) 1UCS



(b) 1BBA



(c) 2PNE

Slika 7.4 Poravnanje originalne i predviđene strukture za proteine 1UCS, 1BBA i 2PNE

Poglavlje 8

Zaključak

AlphaFold 2 pokazao se kao značajan korak naprijed u rješavanju problema savijanja proteina. Tvrdnje nekih da je izlaskom AlphaFold 2 taj problem potencijalno i riješen pretjerane su s obzirom na nedostatke i ograničenja modela, no to ne umanjuje njegovu važnost i korisnost [12, 13, 54]. Njime se postigao veliki napredak u točnosti i pouzdanosti predviđanja tercijarnih struktura peptidnih lanaca svih duljina, prvi put postižući rezultate usporedive s eksperimentalno određenim strukturama [13, 48]. Usprkos njegovim nedostacima, AlphaFold 2 može znanstvenoj zajednici biti koristan alat za određivanje struktura peptidnih lanaca, makar za stjecanje još jedne perspektive, pogotovo otkako je cijeli njegov izvorni kod dostupan javnosti [69]. Iako nije iznimno računski zahtjevan, nije ni program koji se može pokretati na prosječnom računalu, dijelom zbog zahtjeva za procesorskom snagom i radnom memorijom, a dijelom zbog potrebe za preuzimanjem 2,5 TB proteinskih baza na uređaj, pa je preporučljivo pokretati ga na moćnijim računalima namijenjenim za zahtjevne proračune [8, 13]. Stoga je iznimno korisno što je Deepmind tim u AlphaFold Protein Structure Database objavio strukture preko dvjesto milijuna proteina, pokrivši gotovo cijelu UniProt bazu proteina, smanjivši potrebu za samostalnim pokretanjem AlphaFold 2 modela [70]. Prema EMBL-EBI, 80% struktura iz AlphaFoldove baze dovoljno je točno za većinu primjena, a od toga je 35% svih struktura na razini eksperimentalno određenih struktura [70].

Iako je točnost AlphaFold 2 značajan iskorak glede predviđanja struktura proteina, smatram da je problem savijanja proteina daleko od bivanja riješenim. Model kvalitetno primjenjuje koncepte samopozornosti u *Evoformeru* kako bi što bolje informirao Strukturni modul, što je izvrstan način sabiranja i korištenja dosad poznatih struktura proteina i ispravne primjene tog prijašnjeg znanja, ali u tome i leži dio problema [13]. Model je ograničen na 22 kanonske aminokiseline i ne može predviđati post-translacijske modifikacije peptidnih lanaca, što nije nužno tragedija, jer na većinu poznatih proteina ne utječu ta ograničenja [12, 15]. Dakle, AlphaFold 2 je iznimno dobar za predviđanje *poznatog nepoznatog*, strukture peptidnih lanaca koji imaju neke poznate srodnike. Točnost predviđanja drastično opada čim nema dovoljno mnogo povezanih sekvenci kojima model informira svoje zaključivanje strukture. Zbog toga što toliko ovisi o poznatim strukturama, AlphaFold 2 nije toliko prikladan za predviđanje *nepoznatog nepoznatog*, onih struktura o kojima ne znamo gotovo ništa [12]. No to nije nedostatak isključivo AlphaFolda, već svih alata za predviđanje tercijarne strukture proteina koji se temelje na primjeni umjetne inteligencije, koji nepoznatu strukturu predviđaju pokušavajući primijeniti poznate strukture [13]. No, da bi se moglo ispravno modelirati savijanje proteina, potrebno je prvo shvatiti na koji se način odvija proces savijanja proteina u prirodi, pa na temelju toga računalno simulirati fizičke i kemijske zakonitosti koje dirigiraju tercijarnu strukturu peptidnih lanaca. No za takve simulacijske, „racionalne" modele neizbježno je potrebno kudikamo više računalnih resursa u odnosu na „empirijske" modele zasnovane na umjetnoj inteligenciji. Dok se to ne ostvari, AlphaFold 2 je najbolji model za određivanje tercijarne strukture peptidnih lanaca trenutno na raspolaganju.

Literatura

[1] K. A. Dill i J. L. MacCallum, „The Protein-Folding Problem, 50 Years On", *Science*, sv. 338, br. 6110, str. 1042–1046, 2012. DOI: 10.1126/science.1219021. adresa: https://www.science.org/doi/abs/10.1126/science.1219021.

[2] H. Lodish i dr., *Molecular Cell Biology*, 8. izdanje. New York, NY: W.H. Freeman, travanj 2016.

[3] C. A. Orengo, A. E. Todd i J. M. Thornton, „From protein structure to function", *Current Opinion in Structural Biology*, sv. 9, br. 3, str. 374–382, 1999., ISSN: 0959-440X. DOI: https://doi.org/10.1016/S0959-440X(99)80051-7. adresa: https://www.sciencedirect.com/science/article/pii/S0959440X99800517.

[4] C. I. Branden i J. Tooze, *Introduction to Protein Structure*. Garland Science, ožujak 2012.

[5] J. C. Kendrew i dr., „Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å. Resolution", en, *Nature*, sv. 185, br. 4711, str. 422–427, veljača 1960.

[6] K. A. Dill, S. B. Ozkan, M. S. Shell i T. R. Weikl, „The protein folding problem", *Annu. Rev. Biophys.*, sv. 37, br. 1, str. 289–316, 2008.

[7] J. Moult, J. T. Pedersen, R. Judson i K. Fidelis, „A large-scale experiment to assess protein structure prediction methods", *Proteins: Structure, Function, and Bioinformatics*, sv. 23, br. 3, str. ii–iv, 1995. DOI: https://doi.org/10.1002/prot.340230303. adresa: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.340230303.

[8] J. Jumper i dr., „Highly accurate protein structure prediction with AlphaFold", en, *Nature*, sv. 596, br. 7873, str. 583–589, kolovoz 2021.

[9] J. Jumper i dr., „Applying and improving AlphaFold at CASP14", *Proteins: Structure, Function, and Bioinformatics*, sv. 89, br. 12, str. 1711–1721, 2021. DOI: https://doi.org/10.1002/prot.26257. adresa: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26257.

[10] J. Skolnick, M. Gao, H. Zhou i S. Singh, „AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function", *Journal of Chemical Information and Modeling*, sv. 61, br. 10, str. 4827–4831, rujan 2021. DOI: 10.1021/acs.jcim.1c01114. adresa: https://doi.org/10.1021/acs.jcim.1c01114.

[11] K. M. Ruff i R. V. Pappu, „AlphaFold and Implications for Intrinsically Disordered Proteins", *Journal of Molecular Biology*, sv. 433, br. 20, str. 167 208, 2021., From Protein Sequence to Structure at Warp Speed: How Alphafold Impacts Biology, ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2021.167208. adresa: https://www.sciencedirect.com/science/article/pii/S0022283621004411.

[12] S. Robinson, „Artificial intelligence for microbial biotechnology: beyond the hype", *Microbial biotechnology*, sv. 15, listopad 2021. DOI: 10.1111/1751-7915.13943.

[13] C. Outeiral Rubiera. „AlphaFold 2 is here: what's behind the structure prediction miracle". (srpanj 2021.), adresa: https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/ (pogledano 8. 8. 2022.).

[14] W. R. Pearson, „Using the FASTA Program to Search Protein and DNA Sequence Databases", *Computer Analysis of Sequence Data: Part I*, A. M. Griffin i H. G. Griffin, ur. Totowa, NJ: Humana Press, 1994., str. 307–331, ISBN: 978-1-59259-511-2. DOI: 10.1385/0-89603-246-9:307. adresa: https://doi.org/10.1385/0-89603-246-9:307.

[15] A. Ambrogelly, S. Palioura i D. Söll, „Natural expansion of the genetic code", *Nature Chemical Biology*, sv. 3, br. 1, str. 29–35, prosinac 2006. DOI: 10.1038/nchembio847. adresa: https://doi.org/10.1038/nchembio847.

[16] „The DDBJ/ENA/GenBank Feature Table Definition". (studeni 2021.), adresa: https://www.insdc.org/submitting-standards/feature-table/#7.4.3 (pogledano 2. 8. 2022.).

[17] K. Li i dr., „ANDES: Statistical tools for the ANalyses of DEep Sequencing", *BMC research notes*, sv. 3, br. 1, str. 1–12, 2010. DOI: 10.1186/1756-0500-3-199. adresa: https://doi.org/10.1186/1756-0500-3-199.

[18] D. S. Marks i dr., „Protein 3D Structure Computed from Evolutionary Sequence Variation", *PLOS ONE*, sv. 6, br. 12, str. 1–20, prosinac 2011. DOI: 10.1371/journal.pone.0028766. adresa: https://doi.org/10.1371/journal.pone.0028766.

[19] B. Adhikari, „A fully open-source framework for deep learning protein real-valued distances", *Scientific Reports*, sv. 10, br. 1, kolovoz 2020. DOI: 10.1038/s41598-020-70181-0. adresa: https://doi.org/10.1038/s41598-020-70181-0.

[20] R. Shrestha i dr., „Assessing the accuracy of contact predictions in CASP13", *Proteins: Structure, Function, and Bioinformatics*, sv. 87, br. 12, str. 1058–1068, 2019. DOI: https://doi.org/10.1002/prot.25819. adresa: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25819.

[21] L. S. Johnson, S. R. Eddy i E. Portugaly, „Hidden Markov model speed heuristic and iterative HMM search procedure", *BMC Bioinformatics*, sv. 11, br. 1, kolovoz 2010. DOI: 10.1186/1471-2105-11-431. adresa: https://doi.org/10.1186/1471-2105-11-431.

[22] M. Remmert, A. Biegert, A. Hauser i J. Söding, „HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment", *Nature Methods*, sv. 9, br. 2, str. 173–175, prosinac 2011. DOI: 10.1038/nmeth.1818. adresa: https://doi.org/10.1038/nmeth.1818.

[23] M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger i J. Söding, „HH-suite3 for fast remote homology detection and deep protein annotation", *BMC Bioinformatics*, sv. 20, br. 1, rujan 2019. DOI: 10.1186/s12859-019-3019-7. adresa: https://doi.org/10.1186/s12859-019-3019-7.

[24] J. Soding, „Protein homology detection by HMM-HMM comparison", *Bioinformatics*, sv. 21, br. 7, str. 951–960, studeni 2004. DOI: 10.1093/bioinformatics/bti125. adresa: https://doi.org/10.1093/bioinformatics/bti125.

[25] S. Russell i P. Norvig, *Artificial intelligence*, 4. izdanje. Upper Saddle River, NJ: Pearson, studeni 2020. adresa: http://aima.cs.berkeley.edu/.

[26] A. Krogh, M. Brown, I. Mian, K. Sjölander i D. Haussler, „Hidden Markov Models in Computational Biology: Applications to Protein Modeling", *Journal of Molecular Biology*, sv. 235, br. 5, str. 1501–1531, 1994., ISSN: 0022-2836. DOI: 10.1006/jmbi.1994.1104. adresa: https://www.sciencedirect.com/science/article/pii/S0022283684711041.

[27] L. Rabiner, „A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, sv. 77, br. 2, str. 257–286, 1989. DOI: 10.1109/5.18626.

[28] O. C. Ibe, *Markov processes for stochastic modeling*, en, 2. izdanje, serija Elsevier insights. Philadelphia, PA: Elsevier Science Publishing, lipanj 2013.

[29] S. Eddy, „Profile hidden Markov models", *Bioinformatics (Oxford, England)*, sv. 14, br. 9, str. 755–763, 1998., ISSN: 1367-4803. DOI: 10.1093/bioinformatics/14.9.755. adresa: https://doi.org/10.1093/bioinformatics/14.9.755.

[30] „What are profile hidden Markov models?" (2018.), adresa: https://www.ebi.ac.uk/training/online/courses/pfam-creating-protein-families/what-are-profile-hidden-markov-models-hmms/ (pogledano 12. 5. 2022.).

[31] A. L. Mitchell i dr., „MGnify: the microbiome analysis resource in 2020", *Nucleic Acids Research*, sv. 48, br. D1, str. D570–D578, studeni 2019., ISSN: 0305-1048. DOI: 10.1093/nar/gkz1035. adresa: https://doi.org/10.1093/nar/gkz1035.

[32] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu i the UniProt Consortium, „UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches", *Bioinformatics*, sv. 31, br. 6, str. 926–932, studeni 2014., ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu739. adresa: https://doi.org/10.1093/bioinformatics/btu739.

[33] M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding i M. Steinegger, „Uniclust databases of clustered and deeply annotated protein sequences and alignments", *Nucleic Acids Research*, sv. 45, br. D1, str. D170–D176, studeni 2016., ISSN: 0305-1048. DOI: 10.1093/nar/gkw1081. adresa: https://doi.org/10.1093/nar/gkw1081.

[34] M. Steinegger, M. Mirdita i J. Söding, „Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold", en, *Nat. Methods*, sv. 16, br. 7, str. 603–606, srpanj 2019.

[35] J. D. Westbrook i dr., „PDBx/mmCIF Ecosystem: Foundational Semantic Tools for Structural Biology", *Journal of Molecular Biology*, sv. 434, br. 11, str. 167 599, 2022., Computation Resources for Molecular Biology, ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2022.167599. adresa: https://www.sciencedirect.com/science/article/pii/S0022283622001796.

[36] P. E. Bourne, H. M. Berman, B. McMahon, K. D. Watenpaugh, J. D. Westbrook i P. M. Fitzgerald, „[30] Macromolecular crystallographic information file", *Macromolecular Crystallography Part B*, serija Methods in Enzymology, sv. 277, Academic Press, 1997., str. 571–590. DOI: https://doi.org/10.1016/S0076-6879(97)77032-0. adresa: https://www.sciencedirect.com/science/article/pii/S0076687997770320.

[37] T. Lassmann, O. Frings i E. L. L. Sonnhammer, „Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features", *Nucleic Acids Research*, sv. 37, br. 3, str. 858–865, prosinac 2008., ISSN: 0305-1048. DOI: 10.1093/nar/gkn1006. adresa: https://doi.org/10.1093/nar/gkn1006.

[38] sveti Jeronim Stridonski, *Biblia Sacra Iuxta Vulgatam Versionem*, la, 5. izdanje, R. Weber i R. Gryson, ur. Stuttgart, Njemačka: Deutsche Bibelgesellschaft, prosinac 1990.

[39] M. Mohammadi-Kambs, K. Hölz, M. M. Somoza i A. Ott, „Hamming Distance as a Concept in DNA Molecular Recognition", *ACS Omega*, sv. 2, br. 4, str. 1302–1308, travanj 2017. DOI: 10.1021/acsomega.7b00053. adresa: https://doi.org/10.1021/acsomega.7b00053.

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever i R. Salakhutdinov, „Dropout: A Simple Way to Prevent Neural Networks from Overfitting", *Journal of Machine Learning Research*, sv. 15, br. 56, str. 1929–1958, 2014. adresa: http://jmlr.org/papers/v15/srivastava14a.html.

[41] B. A. Rosenfeld, *A history of non-euclidean geometry*, en, 1988. izdanje, serija Studies in the History of Mathematics and Physical Sciences. New York, NY: Springer, rujan 1988.

[42] S. Zhang, Y. Tay, L. Yao i Q. Liu, „Quaternion Knowledge Graph Embeddings", *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox i R. Garnett, ur., sv. 32, Curran Associates, Inc., 2019. adresa: https://proceedings.neurips.cc/paper/2019/file/d961e9f236177d65d21100592edb0769-Paper.pdf.

[43] C. Baldauf i M. Rossi, „Going clean: Structure and dynamics of peptides in the gas phase and paths to solvation", *Journal of physics. Condensed matter : an Institute of Physics journal*, sv. 27, str. 493 002, studeni 2015. DOI: 10.1088/0953-8984/27/49/493002.

[44] „Side chain conformation". (travanj 1995.), adresa: https://www.cryst.bbk.ac.uk/PPS95/course/3_geometry/conform.html (pogledano 12. 8. 2022.).

[45] K. He, X. Zhang, S. Ren i J. Sun, „Deep Residual Learning for Image Recognition", *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016., str. 770–778. DOI: 10.1109/CVPR.2016.90.

[46] P. Eastman i dr., „OpenMM 7: Rapid development of high performance algorithms for molecular dynamics", *PLOS Computational Biology*, sv. 13, br. 7, str. 1–17, srpanj 2017. DOI: 10.1371/journal.pcbi.1005659. adresa: https://doi.org/10.1371/journal.pcbi.1005659.

[47] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg i C. Simmerling, „Comparison of multiple Amber force fields and development of improved protein backbone parameters", *Proteins: Structure, Function, and Bioinformatics*, sv. 65, br. 3, str. 712–725, 2006. DOI: https://doi.org/10.1002/prot.21123. adresa: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21123.

[48] J. Jumper i dr., „AlphaFold 2", *In Fourteenth Critical Assessment of Techniques for Protein Structure Prediction*, 2020.

[49] V. N. Maiorov i G. M. Crippen, „Significance of Root-Mean-Square Deviation in Comparing Three-dimensional Structures of Globular Proteins", *Journal of Molecular Biology*, sv. 235, br. 2, str. 625–634, 1994., ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.1994.1017. adresa: https://www.sciencedirect.com/science/article/pii/S0022283684710175.

[50] A. Zemla, „LGA: A method for finding 3D similarities in protein structures", en, *Nucleic Acids Res.*, sv. 31, br. 13, str. 3370–3374, srpanj 2003. DOI: https://doi.org/10.1093/nar/gkg571.

[51] W. Li, R. D. Schaeffer, Z. Otwinowski i N. V. Grishin, „Estimation of Uncertainties in the Global Distance Test (GDT_TS) for CASP Models", *PLOS ONE*, sv. 11, br. 5, str. 1–16, svibanj 2016. DOI: 10.1371/journal.pone.0154786. adresa: https://doi.org/10.1371/journal.pone.0154786.

[52] V. Mariani, M. Biasini, A. Barbato i T. Schwede, „lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests", *Bioinformatics*, sv. 29, br. 21, str. 2722–2728, kolovoz 2013., ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt473. adresa: https://doi.org/10.1093/bioinformatics/btt473.

[53] A.-M. Šimundić, „Interval pouzdanosti", *Biochemia Medica*, sv. 18, br. 2, str. 154–161, 2008. adresa: https://hrcak.srce.hr/24138.

[54] „Alphafold: A solution to a 50-year-old Grand Challenge in Biology". (studeni 2020.), adresa: https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology (pogledano 4. 8. 2022.).

[55] K. Tunyasuvunakool i dr., „Highly accurate protein structure prediction for the human proteome", *Nature*, sv. 596, br. 7873, str. 590–596, 2021.

[56] M. Necci, D. Piovesan i S. C. Tosatto, „Critical assessment of protein intrinsic disorder prediction", *Nature methods*, sv. 18, br. 5, str. 472–481, 2021.

[57] P. E. Wright i H. Dyson, „Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm", *Journal of Molecular Biology*, sv. 293, br. 2, str. 321–331, 1999., ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.1999.3110. adresa: https://www.sciencedirect.com/science/article/pii/S0022283699931108.

[58] „Povijest Klastera Isabelle". (ožujak 2021.), adresa: https://wiki.srce.hr/display/RKI/Povijest+klastera+Isabelle (pogledano 11. 9. 2022.).

[59] „Tehničke specifikacije Klastera Isabelle". (svibanj 2020.), adresa: https://wiki.srce.hr/pages/viewpage.action?pageId=49284364 (pogledano 11. 9. 2022.).

[60] „Alphafold 2 na Računalnom klasteru Isabella". (veljača 2022.), adresa: https://wiki.srce.hr/display/RKI/Alphafold2 (pogledano 11. 9. 2022.).

[61] C. Zardecki, S. Dutta, D. S. Goodsell, M. Voigt i S. K. Burley, „RCSB Protein Data Bank: A Resource for Chemical, Biochemical, and Structural Explorations of Large and Small Biomolecules", *Journal of Chemical Education*, sv. 93, br. 3, str. 569–575, 2016. DOI: 10.1021/acs.jchemed.5b00404. adresa: https://doi.org/10.1021/acs.jchemed.5b00404.

[62] X. Li, M. Sutcliffe, T. Schwartz i C. Dobson, *Sequence-specific proton NMR assignments and solution structure of bovine pancreatic polypeptide*, listopad 1993. DOI: 10.2210/pdb1bba/pdb. adresa: https://doi.org/10.2210/pdb1bba/pdb.

[63] T.-P. Ko, H. Robinson, Y.-G. Gao, C.-H. Cheng, A. DeVries i A.-J. Wang, *Type III Antifreeze Protein RD1 from an Antarctic Eel Pout*, svibanj 2003. DOI: 10.2210/pdb1ucs/pdb. adresa: https://doi.org/10.2210/pdb1ucs/pdb.

[64] M. Kamo, N. Kudo, W. Lee, K. Ito, H. Motoshim i M. T. and, *The crystal structure of peptide deformylase from Thermus thermophilus HB8*, prosinac 2004. DOI: 10.2210/pdb1v3y/pdb. adresa: https://doi.org/10.2210/pdb1v3y/pdb.

[65] E. Kim, K.-H. Kim, J. Moon, K. Choi, H. Lee i H. Park, *Structures of actinonin bound peptide deformylases from E. faecalis and S. pyogenes*, ožujak 2008. DOI: 10.2210/pdb2os3/pdb. adresa: https://doi.org/10.2210/pdb2os3/pdb.

[66] A. Schmidt, M. Teeter, E. Weckert i V. Lamzin, *Crystal structure of small protein crambin at 0.48 A resolution*, svibanj 2011. DOI: 10.2210/pdb3nir/pdb. adresa: https://doi.org/10.2210/pdb3nir/pdb.

[67] Z. Dauter, D. Liebschner, M. Dauter i A. Brzuszkiewicz, *Bovine trypsin at 0.75 resolution*, prosinac 2012. DOI: 10.2210/pdb4i8h/pdb. adresa: https://doi.org/10.2210/pdb4i8h/pdb.

[68] E. F. Pettersen i dr., „UCSF ChimeraX: Structure visualization for researchers, educators, and developers", *Protein Science*, sv. 30, br. 1, str. 70–82, 2021. DOI: https://doi.org/10.1002/pro.3943. adresa: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3943.

[69] *Open source code for alphafold.* adresa: https://github.com/deepmind/alphafold (pogledano 20. 7. 2022.).

[70] E. Callaway, „'The entire protein universe': AI predicts shape of nearly every known protein", *Nature*, sv. 608, br. 7921, str. 15–16, srpanj 2022. DOI: 10.1038/d41586-022-02083-2. adresa: https://doi.org/10.1038/d41586-022-02083-2.

[71] A. Pribanić, „Reprezentativno učenje u obradi prirodnog jezika", Završni rad, University of Zagreb. Faculty of Humanities, Social Sciences. Department of information i communication sciences, 2021. adresa: https://urn.nsk.hr/urn:nbn:hr:131:538908.

[72] A. Vaswani i dr., „Attention is All you Need", *Advances in Neural Information Processing Systems*, I. Guyon i dr., ur., sv. 30, Curran Associates, Inc., 2017. adresa: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Pojmovnik

GDT Global Distance Test. 30, 31, 35

HMM Hidden Markov Model. 9

IPA Invariant Point Attention. 27

lDDT Local Distance Difference Test. 30, 31, 33, 35, 36

MLP Multi-layer Perceptron. 23

MSA Multiple Sequence Alignment. 4–6, 9, 11, 13, 14, 16, 17, 19–25, 29, 33, 35, 52–54

PHMM Profile Hidden Markov Model. 11–13, 17

RNN Recurrent Neural Network. 29

transformator eng. *transformer*, neuronska mreža temeljena na konceptima samopozornosti (eng. *self-attention*) i kodiranja-dekodiranja [71, 72]. 3, 16, 20

Sažetak

U ovom radu predstavljen je princip rada programskog alata AlphaFold 2, s objašnjenjima primjene genetskog i strukturnog pretraživanja proteina, *Evoformer* transformatora koji izvlači znanje iz reprezentacija MSA i parovima te strukturnog modula koji predviđa trodimenzionalne koordinate svih atoma strukture na temelju reprezentacija iz *Evoformera*. Objašnjene su i metrike RMSD, GDT i lDDT koje služe za evaluiranje sličnosti struktura. Dodatno, praktično je istražena mogućnost predviđanja tercijarne strukture za odabrani skup sekvenci duljina između 36 i 223 aminokiseline korištenjem računalnog klastera Isabella. Dobivene strukture imale su pouzdanost između 19,21 i 99,61 GDT, s vremenima izvršavanja između 35 i 53 minute.

Ključne riječi — **AlphaFold 2, predviđanje strukture proteina, peptidi**

Abstract

This paper presents the core principles of the AlphaFold 2 model, with explanations of genetic and structural protein search, the *Evoformer* transformer used for extracting knowledge from MSA and pair representations, and the Structure module which predicts 3D coordinates of all atoms based on the *Evoformer's* representations. Structure similarity metrics RMSD, GDT and lDDT are also explained. Additionally, the viability of tertiary structure prediction for a select set of sequences between 36 and 223 amino acids is explored using the Isabella computer cluster. The GDT scores of the resulting structures range between 19.21 and 99.61, with computing times between 35 and 53 minutes.

Keywords — **AlphaFold 2, protein structure prediciton, peptides**

Dodatak A

Popis obilježja korištenih u Evoformeru

Dimenzije korištene u Tablici A.1:

- N_{res} - broj rezidua
- N_{clust} - broj grupa MSA
- N_{extra_seq} - broj sekvenci koje nisu ni u jednoj grupi MSA
- N_{templ} - broj strukturnih predložaka

Tablica A.1 Obilježja korištena u Evoformeru

Obilježje Oblik	Opis
aatype $[N_{res}, 21]$	One-hot prikaz ulazne sekvence (21 predstavlja simbole dvadeset aminokiselina + jedan za nepoznato)
cluster_msa $[N_{clust}, N_{res}, 23]$	One-hot prikaz središta grupa MSA (23 predstavlja simbole dvadeset aminokiselina + nepoznanica + razmak + *msa_masked_token*)
cluster_has_deletion $[N_{clust}, N_{res}, 1]$	Binarna vrijednost koja kazuje postoji li brisanje lijevo od rezidue u središtu grupe MSA
cluster_deletion_value $[N_{clust}, N_{res}, 1]$	Količina brisanja s lijeve strane za svaku reziduu u središtu grupe MSA, normalizirano na interval [0, 1] koristeći $\frac{2}{\pi} arctg \frac{d}{3}$ gdje d predstavlja cijeli broj brisanja
cluster_deletion_mean $[N_{clust}, N_{res}, 1]$	Prosječna količina brisanja za svaku reziduu u središtu grupe MSA, računa se kao $\frac{1}{n}\sum_{i=1}^{n} d_{ij}$ gdje je n broj sekvenci u grupi a d_{ij} broj brisanja lijevo od j-te rezidue i-te sekvence. To se potom normalizira koristeći $\frac{2}{\pi} arctg \frac{\overline{d}}{3}$ gdje $\overline{d}$ predstavlja prosječan broj brisanja
cluster_profile $[N_{clust}, N_{res}, 23]$	Distribucija aminokiselina za svaku reziduu u svakoj grupe MSA (23 predstavlja simbole dvadeset aminokiselina + nepoznanica + razmak + *msa_masked_token*)

Obilježje Oblik	Opis
extra_msa $[N_{extra\_seq}, N_{res}, 23]$	One-hot prikaz sekvenci koje nisu središta grupa MSA (23 predstavlja simbole dvadeset aminokiselina + nepoznanica + razmak + *msa_masked_token*)
extra_msa_has_deletion $[N_{extra\_seq}, N_{res}, 1]$	Binarna vrijednost koja kazuje postoji li brisanje lijevo od rezidue iz extra MSA
extra_msa_deletion_value $[N_{extra\_seq}, N_{res}, 1]$	Količina brisanja s lijeve strane za svaku reziduu iz extra MSA, normalizirana na interval [0, 1] koristeći $\frac{2}{\pi} arctg \frac{d}{3}$ gdje d predstavlja cijeli broj brisanja
template_aatype $[N_{templ}, N_{res}, 22]$	One-hot prikaz ulazne sekvence (22 predstavlja simbole dvadeset aminokiselina + nepoznato + razmak)
template_mask $[N_{templ}, N_{res}]$	Maska koja prikazuje postoji li rezidua iz predloška te ima li koordinate
template_pseudo_beta_mask $[N_{templ}, N_{res}]$	Maska koja prikazuje ima li beta-ugljikov atom[1] koordinate u predlošku za određenu reziduu
template_backbone_frame_mask $[N_{templ}, N_{res}]$	Maska koja prikazuje postoje li u predlošku koordinate svih atoma potrebnih za izračun okosnice
template_distogram $[N_{templ}, N_{res}, N_{res}, 39]$	One-hot upareno obilježje koje prikazuje udaljenosti između beta-ugljikovih atoma[1]. Udaljenosti se grupiraju u 38 kategorija jednakih širina između 3,25 Å i 50,75 Å te dodatna kategorija za udaljenosti veće od 50,75 Å

[1]U slučaju glicina, gledaju se alfa-ugljikovi atomi

Obilježje Oblik	Opis
template_unit_vector [N_{templ}, N_{res}, N_{res}, 3]	Jedinični vektor pomaka alfa-ugljikovog atoma unutar lokalnog okvira svake rezidue
template_torsion_angles [N_{templ}, N_{res}, 14]	Tri kuta torzije ovojnice i do četiri kuta torzije bočnih lanaca svake rezidue, prikazane sinusom i kosinusom
template_alt_torsion_angles [N_{templ}, N_{res}, 14]	Alternativni kutovi torzije za dijelove bočnih lanaca sa 180° rotacijskom simetrijom
template_torsion_angles_mask [N_{templ}, N_{res}, 14]	Maska koja prikazuje postoji li u predlošku kut torzije
residue_index [N_{res}]	Indeks rezidue u originalnoj sekvenci