Izbor reprezentativnog skupa membranskih proteina poznate strukture: razvoj poboljšanih algoritama uporabom koncepta nasumičnog modela

Batista, Jadranko

Sveučilište u Splitu
Prirodoslovno-matematički fakultet
Poslijediplomski sveučilišni doktorski studij
Biofizika

Jadranko Batista

IZBOR REPREZENTATIVNOG SKUPA MEMBRANSKIH PROTEINA POZNATE STRUKTURE: RAZVOJ POBOLJŠANIH ALGORITAMA UPORABOM KONCEPTA NASUMIČNOG MODELA

Doktorski rad

Split, 2017.

University of Split
Faculty of Science
Biophysics Doctoral Programme

Jadranko Batista

SELECTION OF REPRESENTATIVE SET OF MEMBRANE PROTEINS OF KNOWN STRUCTURE: DEVELOPMENT OF IMPROVED ALGORITHMS USING THE RANDOM MODEL CONCEPT

Doctoral thesis

Split, 2017.

Sveučilište u Splitu, Prirodoslovno-matematički fakultet

Odjel za fiziku, Poslijediplomski sveučilišni doktorski studij Biofizika

"Izbor reprezentativnog skupa membranskih proteina poznate strukture: razvoj poboljšanih algoritama uporabom koncepta nasumičnog modela"

Doktorski rad autora Jadranka Batiste kao dio obaveza potrebnih za stjecanje doktorata znanosti, izrađen pod vodstvom mentora dr. sc. Bone Lučića, višeg znanstvenog suradnika.

Dobiveni akademski naziv i stupanj: doktor iz područja prirodnih znanosti, polje fizika.

Povjerenstvo za ocjenu doktorskog rada u sastavu:

1. prof. dr. sc. Mile Dželalija, red. prof.
2. prof. dr. sc. Paško Županović, red. prof.
3. dr. sc. Sanja Tomić, znanstvena savjetnica
4. dr. sc. Igor Weber, znanstveni savjetnik, zamjenski član

1. _____

2. _____

3. _____

Potvrđuje da je disertacija obranjena dana 05. 01. 2018.

Voditelj studija: prof. dr. sc. Paško Županović

_____

Predsjednik vijeća studija: prof. dr. dr. h. c. Vlasta Bonačić-Koutecky

_____

ZAHVALE

Najiskrenije zahvaljujem mentoru dr. sc. Boni Lučiću, na istinskom strpljenju, na iscrpnim i stručnim savjetima oko definiranja teme i provedbe istraživanja, na iskazivanju izrazite kolegijalnosti i profesionalnosti te na utrošenom vremenu.

Zahvaljujem svim članovima Vijeća doktorskog studija Biofizike na susretljivošću i razumijevanju i svim profesorima Doktorskog studija Biofizike na prenesenom znanju, te svim kolegama studentima Doktorskog studija Biofizike na ugodnim i korisnim raspravama. Zahvaljujem voditeljima studija prof. dr. sc. Davoru Juretiću i prof. dr. sc. Pašku Županoviću i tajnici studija Ireni Bitunjac na pomoći oko papirologije i na svim pravodobnim informacijama u vezi studija.

Zahvaljujem Ministarstvu znanosti i obrazovanja Republike Hrvatske i Zakladi HAZU (Hrvatska akademija znanosti i umjetnosti, Zagreb, Hrvatska) na financijskoj potpori moga istraživačkoga rada u sklopu izrade disertacije.

Zahvaljujem autorima radova: E. M. Rath, E. Kloppmannu, M. Bernhoferu, B. Rostu, A. Lomizeu, i U. Hobohmu na susretljivosti i pomoći u objašnjavanju njihovih znanstvenih rezlutata (algoritama i baza podataka), te na ugodnoj suradnji. Također, zahvaljujem svim brojnim autorima koji su rješavali strukture membranskih proteina i pohranili ih u proteinsku bazu podataka, te tako omogućili provedbu istraživanja u sklopu izrade moje disertacije. Zahvaljujem kolegi Ivanu Soviću s Instituta Ruđer Bošković iz Zagreba na pomoći u izradi prve programske skripte za obradu dokumenata iz proteinske baze.

Posebno zahvaljujem prof. dr. sc. Davoru Juretiću i akademiku Nenadu Trinajstiću na korisnim savjetima, Zahvaljujem i pokojnom prof. dr. sc. Miljenku Primorcu, na poticaju i ohrabrivanju da se odvažim na istraživanja u području biofizike.

Zahvljujem članovima obitelji: supruzi Slavici na ljubavi i strpljenju, krasnim kćerkama (Loreni, Mariji Sofiji i Elli) na uvijek nasmijanim licima, majci Sofiji na ljubavi i mudrosti, braći i sestrama: Stipi, Milici, Bosiljki, Anti, Zoranu, Nikici, Ivici, Jelenki i Ljubi na svesrdnom bodrenju i potpori.

I na kraju, hvala svim ljudima dobre volje koji tragaju za istinom (u znak sjećanja na pokojnog oca Luku).

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Splitu
Prirodoslovno-matematički fakultet Doktorski rad

IZBOR REPREZENTATIVNOG SKUPA MEMBRANSKIH PROTEINA POZNATE STRUKTURE: RAZVOJ POBOLJŠANIH ALGORITAMA UPORABOM KONCEPTA NASUMIČNOG MODELA

Jadranko Batista

Rad je izrađen u:

- Institutu Ruđer Bošković, Zagreb, Hrvatska i
- Fakultetu prirodoslovno-matematičkih i odgojnih znanosti Sveučilišta u Mostaru, Bosna i Hercegovina.

Sažetak

Strukturu membranskih proteina osjetno je teže eksperimentalno odrediti nego strukturu topljivih proteina. Kako bi se razvio pouzdani model za predviđanje strukture proteina, potrebno je provesti njegovu optimizaciju na što većem (reprezentativnom) skupu membranskih proteina poznatih struktura, međusobnih sličnosti ispod 30%. Postojeći algoritmi za izbor reprezentativnih skupova integralnih membranskih proteina alfa vrste ne koriste informaciju o složenosti strukture, iako se očekuje da će modeli biti pouzdaniji ako su razvijeni na skupu proteina složenijih struktura. Stoga je uveden koncept nasumičnog modela s dvije sekundarne strukture i uočeno da je izraz za procjenu njegove točnosti u vezi sa složenošću strukture. Potom su razvijeni koncepti binomnog i segmentnog nasumičnog modela i izvedeni izrazi za broj mogućih realizacija modelne strukture proteina koji pokazuje analogiju s entropijom. Segmentni nasumični model odgovara strukturi membranskih proteina u kojima više susjednih aminokiselina čini segmente pravilne sekundarne strukture alfa. Broj realizacija modelne strukture segmentnog nasumičnog modela povezan je sa složenošću strukture, i pokazuje značajnu korelaciju s brojem transmembranskih segmenata. To je ugrađeno u razvijene algoritme, a najbolji je temeljen na originalnoj analizi broja zajedničkih susjeda između proteina u početnom skupu. Primjene tih algoritama na baze membranskih proteina poznate strukture daju veće reprezentativne skupove značajno složenijih struktura od onih iz literature.

(176 stranica, 49 slika, 33 tablice, 72 literaturnih navoda, 4 priloga, jezik izvornika: hrvatski jezik)
Rad je pohranjen u:

- Nacionalnoj i sveučilišnoj knjižnici u Zagrebu
- Sveučilišnoj knjižnici u Splitu

Ključne riječi: algoritam, entropija modelne strukture, integralni membranski protein, izbor proteina, matrica sličnosti, nasumični model, segmentni nasumični model, reprezentativni skup, složenost strukture proteina, transmembranski segment

Mentor: dr. sc. Bono Lučić, viši znanstveni suradnik

Ocjenjivači:
1. prof. dr. sc. Mile Dželalija, red. prof.
2. prof. dr. sc. Paško Županović, red. prof.
3. dr. sc. Sanja Tomić, znanstvena savjetnica

Rad prihvaćen 20. 12. 2017.

BASIC DOCUMENTATION CARD

University of Split
Faculty of Science Ph.D. thesis

SELECTION OF REPRESENTATIVE SET OF MEMBRANE PROTEINS OF KNOWN STRUCTURE: DEVELOPMENT OF IMPROVED ALGORITHMS USING THE RANDOM MODEL CONCEPT

Jadranko Batista

Thesis performed at:

- The Ruđer Bošković Institute, Zagreb, Croatia and
- The Faculty of Sciences and Education, University of Mostar, Mostar, Bosnia and Herzegovina

Abstract

It is more difficult to determine experimentally the structure of membrane protein than that of soluble protein. In order to develop a reliable model for predicting protein structure, it is necessary to perform model optimisation on the largest (representative) set of membrane proteins of known structures with mutual similarities below 30%. Existing algorithms for selection of representative sets of membrane proteins of alpha-type do not use information about the complexity of structure, although it is expected that the models will be more reliable in prediction if they are developed on a more complex protein structures. Consequently, the concept of a random model based on two secondary structures was introduced and noticed that the formula for estimation of its accuracy is connected with the complexity of structure. Then, the concepts of the binomial and segmental random model were introduced, as well as formulae for the number of possible realizations of protein model structure, showing the analogy with entropy, were developed. The segmental model is best suited to the membrane protein structure in which several adjacent amino acids form segments having regular secondary structure of alpha type. The number of realizations of model structure of segmental random model is related to the complexity of structure showing a significant correlation with the number of transmembrane segments. It is involved in developed algorithms, and the best one is based on original analysis of the number of common neighbours between proteins in the initial set. Applications of these algorithms to databases of membrane proteins of known structures produce larger representative sets of structures which are significantly more complex than those published in the literature.

(176 pages, 49 figures, 33 tables, 72 references, original in Croatian)
Thesis is deposited in:

- The National and University Library in Zagreb
- University Library in Split.

Keywords: algorithm, model structure entropy, integral membrane protein, selection of proteins, similarity matrix, random model, segmental random model, representative set, protein structure complexity, transmembrane segment

Supervisor: Bono Lučić, Ph.D./Higher research associate

Reviewers:
1. Mile Dželalija, Ph.D./Full professor
2. Paško Županović, Ph.D./Full professor
3. Sanja Tomić, Ph.D./Scientific advisor
Thesis accepted: 20. 12. 2017.

Sadržaj

I. Popis slika

II. Popis tablica

III. Lista simbola i kratica

3D – trodimenzionalni
A1 – Algoritam 1 (algoritam razvijen u disertaciji u analogiji s algoritmom Hobohm 1)
A2 – Algoritam 2 (algoritam razvijen u disertaciji u analogiji s algoritmom Hobohm 2)
A3 – Algoritam 3 (originalni algoritam razvijen u disertaciji)
AK – aminokiselina
AK_{uk} – ukupni broj aminokiselina u reprezentativnom skupu
Bin_{uk} – zbroj vrijednosti binomnog koeficijenta pojedinih lanaca u reprezentativnom skupu
BMRB – Banka podataka bioloških molekula dobivena metodom NMR (engl. *Biological Magnetic Resonance Data Bank*)
d – duljina segmenta
KD – Kyte-Doolittle metoda
N – duljina proteinskog slijeda (duljina slijeda u modelu)
OPM – Baza podataka (i struktura) proteina smještenih u membranu (engl. *Orientations of Proteins in Membrane database*)
PDB – Proteinska baza podataka (engl. *Protein Data Bank*)
PDBe – Europska proteinska baza podataka (engl. *Protein Data Bank in Europe*)
PDBj – Japanska proteinska baza podataka (engl. *Protein Data Bank Japan*)
PDBTM – Baza podataka transmembranskih proteina (engl. *Protein Data Bank of Transmembrane Proteins*)
$Q_{2,ssm}$ – vrijednost koeficijenta Q_2 za cijeli skup po modelu (ne gledaju se proteinski lanci pojedinačno)
$Q_{2,uk}$ – zbroj vrijednosti parametra Q_2 pojedinih lanaca u reprezentativnom skupu
s – broj segmenata
$S_{0,uk}$ – zbroj vrijednosti entropijskog koeficijenta segmentnog nasumičnog modela (bez razmaka između segmenata) pojedinih lanaca u reprezentativnom skupu
$S_{0,sr}$ – srednja vrijednost entropijskog koeficijenta segmentnog nasumičnog modela (bez razmaka između segmenata) u reprezentativnom skupu
TM – transmembranski
TM_{sr} – srednja vrijednost transmembranskih segmenata u reprezentativnom skupu
TM_{uk} – ukupni broj transmembranskih segmenata u reprezentativnom skupu
UniProt – Proteinska baza podataka (engl. *The Universal Protein Resource*)
wwPDB – Svjetska banka za proteinske podatke (engl. *Worldwide Protein Data Bank*)

1. UVOD

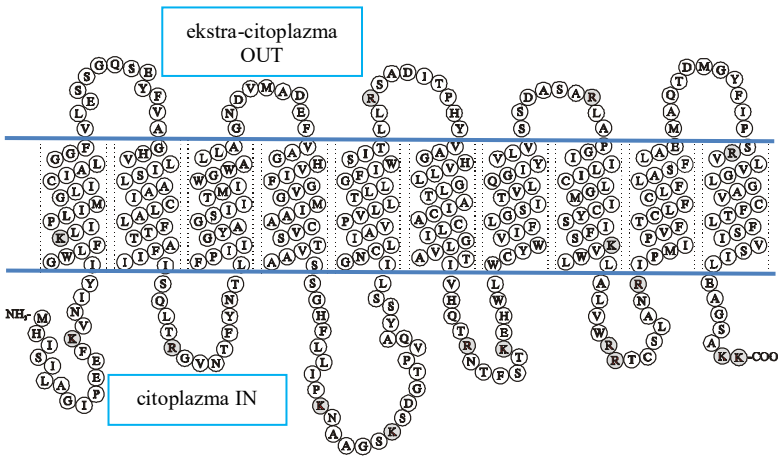
1.1. Osnove glavnih metoda za predviđanje strukture membranskih proteina

Skupina proteina koji svoju funkciju obavljaju dok su u interakciji sa staničnom membranom nazivaju se membranski proteini. Pored ove skupine proteina, koja je (prema procjenama) u svim organizmima kodirana s 20 – 30% kodirajućeg dijela genoma [1], proteine dijelimo još na topljive, vlaknaste i proteine s intrinzično neuređenom trodimenzionalnom (3D) strukturom. Prema položaju u odnosu na membranu i vrsti interakcije s membranom razlikujemo:

- integralne membranske proteine, koji biološku funkciju obavljaju tako što su stalno smješteni (uronjeni) u staničnu membranu, i
- periferne membranske proteine koji obavljaju funkciju privremeno pričvršćeni uz membranu ili drugi integralni membranski protein [2].

Proteini uronjeni u membranu obično jednom ili više puta prolaze kroz nju, usidreni dijelovima koji poprimaju pravilnu sekundarnu strukturu alfa uzvojnice ili sekundarnu strukturu beta. Dijelovi membranskog proteina koji su uronjeni u unutrašnjost membrane nazivaju se još i transmembranski (TM) segmenti. U preostalom dijelu primarne strukture (koji nije u kontaktu s membranom), membranski je protein u kontaktu s unutarstaničnim ili s izvanstaničnim medijem. Prostor u kojemu se nalazi prvi izvan-membranski dio slijeda membranskog proteina (izvan-stanični – 'OUT' ili unutarstanični – 'IN') određuje njegovu topologiju, tj. smjer prolaska preostalog dijela proteinskog slijeda u odnosu na membranu.

Najveća je skupina membranskih proteina ona alfa vrste, u kojoj su TM segmenti u primarnoj strukturi obično dugi 18-22 susjedne aminokiseline i poprimaju sekundarnu strukturu uzvojnice alfa s periodom od 3.5 aminokiseline po jednom zavoju uzvojnice. Na slici 1. shematski je prikazan membranski protein s deset TM segmenata i topologijom određenom N-krajem smještenim u citoplazmi.



Slika 1. Shematski prikaz membranskog proteina s 10 TM segmenata.

Od svih membranskih proteina najviše je poznatih slijedova (primarnih struktura ili lanaca) i 3D struktura membranskih proteina alfa vrste (~70% svih proteinskih lanaca membranskih proteina). Struktura fotosintetskog reakcijskog centra spada u skupinu membranskih proteina alfa vrste, i prva je određena struktura nekog membranskog proteina. Ta je struktura objavljena 1985. godine [3], a za njeno otkriće dodijeljena je Nobelova nagrada za kemiju 1988. godine (dobitnici nagrade: J. Deisenhofer, R. Huber i H. Michel).

Ideje u ranom razvoju metoda za modeliranje i predviđanje sekundarne strukture proteina možda najbolje ilustrira metoda Chou-a i Fasman-a iz 1974. godine [4,5]. Ta metoda predviđa sekundarnu strukturu pojedine aminokiseline u proteinu na temelju specifičnosti rasporeda nekoliko susjednih aminokiselina (u tzv. lokalnoj okolini) u primarnoj strukturi proteina, te na temelju ljestvica (numeričkih vrijednosti) različitih konformacijskih sklonosti aminokiselina za pojedinu vrstu sekundarne strukture. Prve ljestvice konformacijskih sklonosti izračunate su i rabljene u toj metodi na temelju maloga broja proteina kojima je tada bila određena struktura. To su proteini poput inzulina, te više međusobno sličnih primarnih struktura proteina kao što su hemoglobin, mioglobin, lizozim, ribonukleaza, citokrom i dr. Obično se definiraju i modeliraju pravilni dijelovi sekundardne strukture alfa i beta vrste, a sve ostale podvrste svrstavaju se u skupinu nepravilne sekundarne strukture.

Predviđanje sekundarne strukture membranskih proteina alfa vrste započelo je metodom poznatom pod nazivom Kyte-Doolittle (KD) iz 1982. godine [6], te je predviđanje kasnije potpomognuto metodom hidrofobnoga momenta transmembranskih α-uzvojnica [7]. Zanimljivo je spomenuti da je metoda KD razvijena prije nego je riješena prva 3D struktura nekog membranskog proteina. Potom je G. von Heijne razvio metodu za predviđanje topologije membranskih proteina koja je temeljena na: (1) predviđanju položaja hidrofobnih dijelova (TM segmenata) te na (2) analizi nejednolikog rasporeda pozitivno nabijenih aminokiselina (arginina i lizina) u dijelovima slijeda između TM segmenata u bakterijskim proteinima [8,9]. Naime, von Heijne je uočio kako je veći broj pozitivno nabijenih aminokiselina grupiranih s unutarnje strane stanice u dijelovima slijeda koji spajaju susjedne TM segmente.

Za definiranje i optimiranje metode KD, i danas iznimno često citirane u znanstvenoj literaturi (preko 17000 puta), korištene su:

(1) strukturne informacije dobivene s pomoću eksperimentalne metode kojom se, ugradnjom obilježenih proba u proteinskom slijedu, približno određivao položaj i broj TM segmenata u membranskom proteinu, kao i njegova topologija,
(2) eksperimentalni podaci o duljini TM segmenata alfa vrste (usklađene s debljinom membranskoga dvosloja) te
(3) usrednjena hidrofobnost susjednih ~19 aminokiselinskih ostataka u lancu membranskog proteina ($9 + 1 + 9$, odnosno središnja aminokiselina u promatranom dijelu i njoj susjednih devet lijevih i devet desnih aminokiselina; takav dio lanca/slijeda naziva se uobičajeno „prozor širine 19“).

Usrednjene hidrofobnosti za cijeli proteinski slijed računaju se metodom kliznog (pomičnog) prozora za svaku aminokiselinu u aminokiselinskom slijedu promatranu kao središnju u prozoru, i prikazuju se kao profil hidrofobnosti duž proteinskog slijeda. Dakle, svakoj aminokiselini u profilu hidrofobnosti pridružena je prosječna vrijednost hidrofobnosti 9 lijevih, središnje, i 9 desnih susjednih aminokiselina u proteinskom slijedu. Potom su poboljšavane ranije uvedene metode ili razvijane neke novije metode [10], koje su vremenom, u pravilu, postajale sve složenije.

Po jednostavnosti, fizikalnosti i zornosti metodi Kyte-Doolittle najsličnija je metoda (koncept) sklonosnih funkcija [11]. Ta je metoda temeljena na opažanju da 20 prirodnih aminokiselina pokazuje različite sklonosti za lokalne okoline različitih hidrofobnosti. K tome, te se sklonosti razlikuju i ovisno o vrsti sekundarne strukture u kojoj se pojedina aminokiselina pojavljuje, a opisane su empirijskim normalnim raspodjelama za svaku aminokiselinu i za svaku vrstu sekundarne strukture. Osnovni koncept sklonosnih funkcija potom je nadograđen potrebnim podprogramima za glađenje sklonosnih funkcija, prepoznavanje specifičnih motiva u predviđanjima i modifikaciju metode kliznoga prozora na rubovima aminokiselinskog slijeda. Nadalje, u metodi sklonosnih funkcija razvijen je i podprogram za analizu i izračun parametara za iskazivanje i procjenu točnosti i pouzdanosti na skupu za učenje, i na vanjskom skupu proteinskih slijedova. Tako nadograđen i optimiran koncept sklonosnih funkcija uobličen je u

metodu nazvanu SPLIT [12,13], koja je kasnije postavljena na internetski poslužitelj [14]. Naposljetku, metoda SPLIT nadograđena je i analizom raspodjele nabijenih aminokiselina u slijedu kako bi se mogla predvidjeti i topologija membranskoga proteina [15].

Razvoj metoda za predviđanje sekundarne strukture proteina ubrzava se najviše radom grupe Chrisa Sandera (Heidelberg) nakon 1990. godine, i ponajprije zahvaljujući:

(a) povećanju broja proteina kojima je riješena struktura,
(b) prepoznavanju važnosti evolucijske informacije kroz povećanu očuvanost dijelova primarne strukture koji poprimaju pravilnu sekundarnu strukturu (uglavnom alfa i beta vrste) i njenim uključivanjem u postupak optimiranja (učenja) metode, te
(c) uvođenjem u uporabu u području modeliranja sekundarne strukture proteina nelineranih i složenih metoda neuronskih mreža [16,17].

Kasnije se razvijaju i uvode brojne metode koje su unaprijeđene ponajviše u smislu uporabe složenijih algoritama, ansambla više različitih metoda, i kroz detaljnije optimiranje parametara modela. Možda ponajbolji primjer takvih metoda razvijen za topljive proteine metoda je Jpred koja je optimirani ansambl (skup) više metoda [18,19], i postiže prosječnu točnost predviđanja alfa, beta i nepravilne sekundarne strukture od preko 84%.

1.2. Problem izbora reprezentativnog skupa proteina niske međusobne sličnosti

Dodatni bitan čimbenik koji je značajno doprinio povećanju točnosti metoda za predviđanje sekundarne strukture proteina odnosi se na prepoznavanje potrebe izbora (iz cjelokupne proteinske baze podataka (PDB) [20]) reprezentativnog skupa proteina niske zalihosti (tj. sličnosti < 25%) [21] u primarnoj strukturi proteina. Takav skup potreban je u razvoju metoda kako bi algoritmi bili optimirani (bili naučeni) na primjerima (primarnim strukturama) koji se čim više razlikuju, te tako bila u stanju pronaći dovoljno općenita a ipak specifična i statistički značajna pravila lokalnog organiziranja aminokiselina u primarnoj strukturi koja određuju (uvjetuju) sekundarnu strukturu proteina. U protivnom, metode optimirane na skupu proteina (skupu 'za učenje' metode) visoke sličnosti u primarnim strukturama (koji imaju i slične sekundarne i 3D strukture) bile bi u stanju pouzdano predviđati sekundarnu strukturu samo onih proteina koji su dovoljno slični proteinima iz skupa za učenje.

Analizom primarnih i 3D struktura tada poznatih proteina uočeno je da, ukoliko je sličnost dvaju proteinskih slijedova u primarnoj strukturi ispod 25% (odnosno identičnost ispod 20%), tada nema sličnosti (ne prenosi se sličnost) na 3D strukturu proteina [21]. Prema tome, za sve proteinske slijedove koje imaju međusobno sličnost ispod tog praga kaže se da imaju originalne (jedinstvene) primarne i 3D strukture, pa time i originalne sekundarne strukture. U tim se analizama sličnost između proteinskih lanaca analizira nekom od brojnih realizacija algoritma BLAST [22,23]. Uvedene su dvije inačice algoritma za izbor reprezentativnog skupa proteina nazvane u literaturi po prvom autoru kao Hobohm 1 i Hobohm 2 [24,25], koje su i danas često rabljeni algoritmi. Ti su algoritmi povezani s mrežnom inačicom proteinske baze podataka tako da ih zainteresirani korisnici mogu koristiti za izdvajanje podskupa proteinskih slijedova čija je međusobna sličnost manja od unaprijed definiranoga praga [20].

Uporabom tih algoritama [24,25] za izbor reprezentativnog skupa proteina u radu kojem su značajno unaprijedili predviđanje sekundarne strukture proteina [16], autori su dobili 1993. godine 130 jedinstvenih proteinskih lanaca. Njihove primarne strukture (skoro svi topljivi proteini uz samo tri membranska) bile su slične manje od 30%, te pri tom nije uočena značajna sličnost između njihovih 3D struktura. Danas je taj broj jedinstvenih lanaca topljivih proteina narastao preko 5000. Svi ti jedinstveni proteinski lanci svrstavaju se u manje od 1300 (prema bazi CATH [26]) odnosno 1400 (prema bazi SCOP [27]) strukturnih oblika (engl. fold ili folding types), a zanimljiv je podatak da je posljednjih godina vrlo usporen porast broja novih

strukturnih oblika [27]. Iako je broj novih proteinskih struktura koje se pohranjuju svake godine u bazu PDB u stalnom (velikom) porastu, ovaj podatak o (skoro) zaustavljenom porastu broja novih strukturnih oblika pokazuje kako je vjerojatno iscrpljena 'originalnost' struktura proteina (gledano prema broju strukturnih oblika, a i po sličnosti 3D strukture kao i sekundarne strukture proteina). Naravno, pritom se misli samo na proteine izolirane iz postojećih živih organizama u prirodi, a ne na one umjetno stvorene.

Dok je broj topljivih proteina riješene strukture rastao relativno brzo, broj membranskih proteina kojima je određena struktura bitno je manji. Iako se procjenjuje da je broj membranskih proteina kodiran genima u svim genomima (u prosjeku) 20 - 30% [1], od ukupno 120.000 proteinskih struktura pohranjenih u PDB [20], membranskih proteina svega je 3380, s ukupno 6790 proteinskih lanaca [28,29].

1.3. Motivacija za istraživanje (razvoj algoritama za izbor reprezentativnog skupa membranskih proteina)

Problem velikog nesrazmjera u broju određenih struktura membranskih i topljivih proteina bio je stalno prisutan u proteklih 32 godine (od prve objavljene riješene strukture membranskoga proteina 1985. godine [3]). Iz toga zaključujemo da u optimizaciji brojnih metoda razvijenih u svrhu predviđanja i modeliranja strukture membranskih proteina u prošlosti koristilo *djelomično točne strukture*. Takve su strukture bile (tada) pohranjene u bazi SWISS-PROT [30,31], a danas su uključene u bazu UniProt [32]. Za utvrđivanje (određivanje, definiranje) sekundarne strukture i položaja TM segmenata proteina u tom periodu korišten je (u pravilu) i ranije spomenuti algoritam Kyte-Doolittle [6]. Stoga možemo reći da su takve strukture modelne, a ne eksperimentalno određene strukture. Zbog toga su, strogo uzeto, sve ranije razvijene i danas najčešće korištene metode za predviđanje strukture membranskih proteina, koje su u postupku optimiranja (učenja) koristile te približno točne tj. modelne strukture, zapravo približno točne metode [33-39].

Početkom istraživanja u sklopu doktorske disertacije uočen je taj problem, i odlučeno je najprije pronaći (izabrati) najveći skup reprezentativnih membranskih proteina poznate strukture. Potom je bilo planirano iskoristiti taj skup kako bi se na njemu provjerila točnost razvijenih i dostupnih metoda u predviđanju strukture proteina. Nadalje, bilo je planirano iskoristiti takav skup u provjeri točnosti, te u optimizaciji i nadogradnji metoda poput metode SPLIT [11,15] i metode KD [6]. Preliminarni rezultati, u kojima je u skupu dobiveno 143 jedinstvenih proteinskih lanaca (s međusobnim sličnostima u primarnim strukturama manjim od 30%), prezentirani su u radu na znanstvenome skupu [40]. U to vrijeme pojavljuju se u literaturii radovi drugih istraživačkih grupa u svijetu koje su uočile ovaj problem i na njemu rade, poput grupe Emme Rath (Australia) [41] i grupe Burkharda Rosta (Njemačka) [42,43]. Jedan dio istraživanja planiran u sklopu izrade doktorske disertacije napravljen je u tim radovima s pomoću ranije razvijenih standardnih metoda [25,44]. Kao glavna novost i doprinos u tim radovima bili su veći skupovi i to eksperimentalno određenih struktura proteina, što je posljedica većeg polaznoga skupa riješenih struktura proteina pohranjenih u PDB [20]. Međutim, sve su strukture određene eksperimentalnim postupcima, i na njima se moglo usporediti dostupne metode za predviđanje strukture membranskih proteina [41,42,43]. Uspostavljen je kontakt s tim grupama kako bi se razmjenili istraživački podatatci i usporedili rezultati.

Radeći na istom istraživačkom problemu tijekom preliminarne faze istraživanja planirali smo razviti vlastiti algoritam za izbor reprezentativnog skupa proteina, a ne koristiti algoritam drugih grupa. U samoj pripremi istraživanja analizirali smo strukture membranskih proteina i utvrdili kako je glavni numerički parametar koji određuje složenost strukture i topologije membranskog proteina zapravo broj TM segmenata i duljina proteinskog slijeda. Stoga smo

definirali ovaj jednostavni kriterij broja TM segmenata (koji nije razmatran u ranijim algoritmima [24,25,44]) kao glavni kriterij u odlučivanju *koji protein nastojati uključiti, a koji nastojati izbaciti, iz kasnijih razmatranja i iz reprezentativnog skupa*. Kao drugi (pomoćni) kriterij definirali smo duljinu proteinskog slijeda, koji je djelomično već sadržan u prvom kriteriju stoga što, u pravilu, proteini koji imaju više TM segmenata (segment čine obično 18-22 susjedne aminokiseline) obično imaju i dulji lanac (slijed). S tom novošću u algoritmu, dobiveno je u preliminarnim rezultatima 143 jedinstvenih proteinskih lanaca s međusobnim sličnostima u primarnoj strukturi manjim od 30% [40]. To je bilo osjetno poboljšanje u odnosu na 101 membranski protein koliko je dobiveno u radu grupe Emme Rath pomoću algoritma Hobohm 2 [41,25]. Treba napomenuti kako jedan dio tog poboljšanja dolazi i stoga što se u našem radu [40] polazilo od novijih i većih baza proteinskih struktura. Kad se uzme u obzir doprinos zbog većeg početnog skupa, poboljšanje je iznosilo 10% – 20% gledano po broju proteinskih lanaca. K tome, proteinski lanci izabrani s pomoću našeg (preliminarnog) algoritma sadržavali su, u prosjeku, osjetno veći broj TM segmenata.

Problem izbora reprezentativnog skupa proteinskih lanaca kod kojih je sličnost između svih parova ispod neke unaprijed definirane razine, nije moguće riješiti jednoznačno i optimalno. Za određeni početni skup sastavljen od N proteinskih lanaca potrebno je izračunati međusobnu sličnost (npr. izraženu u postotcima sličnosti u rasponu od $0-100\%$) između svakog para proteina. Tako se u konačnici dobije matrica sličnosti dimenzije $N{\cdot}N$. Potom se definira najveća dopuštena razina sličnosti (tzv. prag sličnosti), tj. najviša postotna sličnost između bilo koja dva para proteina koji ostaju u konačnome skupu. Radi jednostavnosti, sličnosti koje su manje od praga zamijenimo s nulom, a sličnosti iznad tog praga jedinicom. Algoritmom se provodi smanjenje početnoga skupa izuzimanjem po jednoga proteinskog lanca u svakom koraku izvođenja, u nastojanju dobivanja najvećeg mogućeg reprezentativnog skupa sastavljenog od M ($M < N$) lanaca. Međusobne sličnosti između M izabranih proteinskih lanaca manje su od unaprijed definiranog praga sličnosti. Konačni rezultat koji se želi postići po završetku izvođenja algoritma je - što veća jedinična matrica sličnosti dimenzije $M \cdot M$ (gdje je M broj proteina preostalih u konačnome skupu). Taj je problem djelomično sličan problemu putujućeg trgovca u kojem, za unaprijed zadani povezani skup (mrežu) koju čini N cestovno povezanih gradova, treba pronaći optimalni redoslijed prolaženja kroz sve gradove kako bi trgovac, pritom, prešao najmanji mogući put.

Uz unaprijed definirane kriterije po kojima će se u svakom koraku rada algoritma određivati koji od dva proteinska lanca (međusobne sličnosti više od definiranoga praga) izuzimamo a koji u tom koraku zadržavamo u skupu za daljnje analize. Taj postupak izbora jako ovisi o redoslijedu u kojem su proteinski lanci poslagani (sortirani) na početku, što zapravo određuje i redoslijed kojim se analiziraju parovi proteinskih lanaca. Broj mogućih redoslijeda slaganja (sortiranja) N proteinskih lanaca je $N!$. Kako bi se optimiralo i ubrzalo rješavanje problema, potrebno je pronaći neki fizikalno-kemijski (strukturom definirani) kriterij koji će dati manji broj redoslijeda proteina koje je potrebno razmotriti u radu algoritma. Usprkos važnosti tog problema za razvoj novih metoda za predviđanje strukture proteina, pregledom literature uočeno je da se problem razvoja optimalnoga algoritima za izbor reprezentativnog skupa proteina niske sličnosti istraživao u samo tri rada [24,25,44].

1.4. Istraživačka hipoteza i cilj istraživanja

S obzirom na naše preliminarne rezultate izbora reprezentativnog skupa membranskih proteina i njihovu usporedbu s rezultatima dviju istraživačkih grupa [41,42,43], dobivenim s ranije razvijenim algoritmima [25,44], uočeno je da:

(a) za taj problem nije moguće pronaći jednoznačno rješenje,

(b) u literaturi postoje svega dvije vrste algoritama kojima je rješavan taj problem i
(c) da u rješavanju toga problema postojećim algoritmima nisu definirani niti rabljeni kriteriji utemeljeni na analizi samih strukturnih obilježja proteinskog lanca.

Analizom literature i preliminarnih rezultata postavljena je glavna istraživačka hipoteza u provedbi istraživanja u sklopu izrade doktorske disertacije koja glasi: „*Moguće je postići značajan napredak (iskorak) u razvoju algoritama za izbor reprezentativnog skupa membranskih proteina ukoliko se definiraju, uvedu, i koriste u radu algoritama kriteriji složenosti proteinske strukture. Kao jednostavni primjer, u prvoj aproksimaciji, može se rabiti informacija o broju TM segmenata i o duljini proteinskog lanca.*“

Radom na provjeri istraživačke hipoteze definirani su (postavljeni) sljedeći glavni istraživački ciljevi u izradi disertacije:

1) reproducirati algoritme i reproducirati rezultate australske i njemačke istraživačke grupe, te na istim polaznim podacima usporediti rezultate koji će se dobiti uporabom njihovih i naših preliminarnih algoritama [40],
2) analizirajući karakteristična svojstva membranskih proteina pronaći kriterije kako bi se razvio i optimirao originalni algoritam za izbor reprezentativnog skupa membranskih proteina,
3) primijeniti razvijene algoritme na skupove membranskih proteina iz literature i provesti usporedbu s drugim algoritmima, te ih primijeniti na poznate strukture membranskih proteina dostupne u bazi PDB radi izdvajanja reprezentativnih skupova niske međusobne sličnosti,
4) nastojati pronaći fizikalno objašnjenje za definirane kriterije i dobivene rezultate,
5) uspostaviti vezu između kriterija izabranih za iskazivanje kompleksnosti (složenosti) realnih modelnih struktura s kompleksnošću nasumičnih (slučajno generiranih) modelnih struktura, što je jednim dijelom već objavljeno [45,46], i
6) razmotriti mogućnost primjene razvijenih algoritama u drugim srodnim područjima istraživanja.

U provedbi planiranih istraživanja prvi korak je pristup bazama proteinskih podataka poput baza OPM [28,29], PDB [20] i UniProt [32], te preuzimanje, izdvajanje i (stalna) provjera podataka, postupaka i struktura. Glavnina računalnih metoda (algoritmi, skripte, programi) koja se rabi u istraživanjima razvijena je u sklopu izrade disertacije, a realizirane su u programskom jeziku Python [47,48]. U vezi algoritama u literaturi nazvanima Hobohm 1 i 2, uspostavljen je kontakt s jednim od autora tih algoritama (Uwe Hobohm) koji je poslao izvršnu verziju programa algoritma Hobohm 2. Taj je algoritam/program, nakon više iteracija i pokušaja, radio iznimno sporo. Inače, Hobohm 2 najčešće je korišten algoritam (prema znanstvenoj literaturi) za izbor reprezentativnih skupova proteina, i može se reći da je to standardni algoritam u području. Zbog toga su u sklopu izrade disertacije ti algoritmi realizirani u programskom jeziku Python, na temelju opisa algoritama Hobohm 1 i 2 u originalnim radovima [24,25]. Druga metoda koja je objavljena za rješavanje problema izbora reprezentativnog skupa proteina niske zalihosti naziva se UniqueProt [44]. Metodu su razvili S. Mika i B. Rost, a zanimljivo je spomenuti da je bitno manje korištena (i citirana) u znanstvenoj literaturi od metoda Hobohm 1 i 2. Usporedbe razvijenih algoritama u disertaciji s metodom UniqueProt [44] provedene su dijelom u kontaktu s istraživačima grupe profesora Rosta, na više skupova podataka.

Kako bi se provela višestruka usporedba sličnosti među svim parovima proteinskih lanaca u skupu (kojih, nakon preliminarnog pročišćavanja identičnih ili skoro identičnih lanaca, ima više od 1200) instaliran je besplatno dostupan program EMBOSS [49]. Za potrebe preuzimanja, provjere i daljnjeg korištenja izlaznih podataka, te za potrebe priređivanja ulaznih podataka u obliku potrebnom za pokretanje programa EMBOSS, razvijene su potrebne skripte i programi u programskom jeziku Python [47,48].

1.5. Očekivani rezultati

Provedbom planiranih istraživanja u sklopu izrade disertacije očekuje se dobiti znanstveno vrijedne rezultate koji predstavljaju iskorak u odnosu na stanje područja, i to:

1. razviti nove poboljšane algoritme za izbor reprezentativnog skupa membranskih proteina niske zalihosti, uvođenjem fizikalnih kriterija koji kvantificiraju složenost (kompleksnost) strukture membranskih proteina,
2. izvesti izraze za opis složenosti nasumičnih modelnih struktura u slučaju:
   a. kada se promatra struktura na razini aminokiselinskih ostataka u proteinskom lancu, odnosno na razini udjela sekundarnih struktura u proteinu, i
   b. kada se promatra sekundarna struktura α uzvojnice na razini segmenata,
3. na temelju izraza izvedenih za procjenu točnosti i složenosti nasumičnih modela, definirat će se parametri za kvantificiranje složenosti strukture pojedinog lanca odnosno skupa proteina, i dati njihova fizikalna interpretacija,
4. definirani parametri za procjenu složenosti strukture membranskih proteina iz (3) ugradit će se u algoritme za izbor reprezentativnih skupova membranskih proteina iz (1),
5. primjenom poboljšanih algoritama pod (1), izabrat će se novi reprezentativni skupovi membranskih proteina algoritama s više lanaca i više TM segmenata za slučaj kad se postavi prag indentičnosti/sličnosti između proteina. Dobiveni rezultati usporedit će se s rezultatima najčešće korištenih metoda iz literature.

2. MATERIJALI I METODE

U postupku razvoja algoritama za izbor reprezentativnog skupa membranskih proteina iznimno je važno prikupiti što veći skup što preciznije karakteriziranih primarnih, sekundarnih i 3D struktura proteina. Ti se podaci nalaze u proteinskim bazama podataka strukturiranim (organiziranim) po određenim specifičnim pravilima, i dostupnim putem interneta. Najznačajnije proteinske baze podataka, korištene u izdvajanju informacija o proteinskim slijedovima i strukturama potrebne u provedbi istraživanja u disertaciji, bit će opisane u nastavku. Drugi dio metoda (računalnih algoritama i novih koncepata) korištenih za dobivanje, analizu i usporedbu rezultata, po prvi put je razvijen u sklopu izrade ove disertacije i predstavlja sastavni dio rezultata. Stoga su ti algoritmi i teorijski koncepti opisani u poglavlju 3 (Rezultati i rasprava), u dijelu 3.1.

2.1. Proteinske baze podataka

Za analizu membranskih proteina i usporedbu s rezultatima iz literature potrebni su što detaljniji podaci o korištenim metodama i kvaliteti struktura proteina poput eksperimentalne metode pomoću koje je riješena struktura proteina, podatka o rezoluciji, primarnim strukturama proteinskih lanaca, itd. U tu svrhu izrađene su, stalno održavane i nadograđivane proteinske baze podataka među kojima se ističu:

- Worldwide Protein Data Bank (wwPDB) [20,50].
- UniProt [32,51]
- OPM [28,29,52,53]
- PDBTM [54,55,56,57]

Prve dvije baze podataka općenite su (izvorne) baze podataka o proteinskim strukturama, dok su druge dvije baze izrađene specifično za membranske proteine, a sadrže mrežne (internetske) poveznice na izvorne ali i na druge baze podataka koje se odnose na spečifičnosti proteinskih struktura i svojstava.

Za razvoj algoritama koristili su se podaci o položajima TM segmenata iz baze podataka OPM (engl. *Orientations of Proteins in Membranes (OPM) database*) a preuzeti su s internetske stranice baze [28,29]. No, kako slijedovi (primarne strukture) proteinskih lanaca u ovoj bazi imaju prekide, u svrhu prikupljanja i analize cjelovitih proteinskih aminokiselinskih slijedova bilo je potrebno preuzeti cjelokupne slijedove iz proteinske baze podataka PDB [20,50]. U kompletiranju primarne strukture proteina rabljene su, radi provjere, i informacije iz baze UniProt [32,51]. Naime, baza UniProt sadrži informacije o daleko većem broju proteina nego baza PDB, koja sadrži informacije samo o onim proteinima kojima je eksperimentalno određena 3D struktura za dio ili za cijelu primarnu strukturu [20]. Osim baze proteina OPM specijalizirane za membranske proteine [28,29], često se koristi za definiciju strukture membranskih proteina i druga specijalizirana baza pod nazivom PDBTM (engl. *Protein Data Bank of Transmembrane Proteins*) [54,55,56,57].

2.1.1. Worldwide Protein Data Bank (wwPDB)

Proteinska baza podataka PDB [20] globalni je repozitorij informacija o 3D strukturama velikih bioloških molekula, tj. proteina i nukleinskih kiselina. Te su molekule odgovorne za prijenos genetske informacije i održavanje i funkcioniranje živih organizama. Njihove strukture ključne su za funkcioniranje života. U slučaju proteina izoliranih iz ljudskih organizama, proteinske strukture iz baze PDB ključne su za razumijevanje i održavanje ljudskoga zdravlja, a promjene u strukturi i funkciji uzrokom su brojnih bolesti. Analize strukture proteina, nukleinskih kiselina i

njihovih kompleksa, te njihova interakcija s drugim molekulama u osnovi su istraživanja lijekova. Podaci sadržani u ovoj bazi kreću se od struktura malih proteina i dijelova DNA do složenih molekulskih strojeva poput ribosoma i besplatno se mogu koristiti za potrebe istraživanja. Podaci u bazi osvježavaju se na tjednoj bazi.

PDB je osnovana 1971. godine u Nacionalnom laboratoriju u Brookhavenu (engl. *Brookhaven National Laboratory*) pod vodstvom Waltera Hamiltona, i izvorno je sadržavala samo sedam struktura. Zbog uočavanja njene izuzetne važnosti, 1998. godine je za upravljanje PDB-om postala odgovorna organizacija pod nazivom *"Research Collaboratory for Structural Bioinformatics* (RCSB)". Godine 2003. osnovan je wwPDB [50], čija je uloga održavanje jedinstvene arhive makromolekularnih strukturnih podataka koja je slobodno i javno dostupna globalnoj zajednici. Sastoji se od organizacija koje se brinu o prikupljanju, pohranjivanju i obradi podataka i koje su ujedno i centri za distribuciju PDB podataka. Osim navedene baze podataka, sastavnice wwPDB-a još su: PDBe (*Protein Data Bank in Europe*), PDBj (*Protein Data Bank Japan*) i BMRB (b*iological Magnetic Resonance Data Bank*).

Kolika je važnost ovih baza govore i podaci o broju preuzimanja tijekom perioda od jedne godine. Iz tablice 1. vidi se da broj preuzimanja neprestano raste, a najviše ih se ostvari preko adrese RCSB.

Tablica 1. Usporedba broja preuzimanja podataka sa sastavnica wwPDB u 2016. naspram 2010. godine.*

godina	***2016.***		***2010.***	
baza/tip	FTP arhiva	web stranica	FTP arhiva	web stranica
RCSB PDB	**293,648,366**	**161,208,456**	159,248,214	64,569,658
PDBe	30,274,284	44,432,830	34,383,219	14,017,349
PDBj	42,755,247	19,556,904	19,549,533	2,559,003
ukupno	366,677,897	225,198,190	213,180,966	81,146,010
ukupno sve baze	591,876,087		294,326,976	

* = (preuzeto s mrežne stranice baze wwPDB [50], 2017.)

Osim toga, na navedenim mrežnim stranicama korisnici mogu obavljati jednostavne i složene upite u vezi sa strukturnim podatcima (npr. prikupljanje, objedinjavanje i povezivanje), a na raspolaganju su i alati za vizualizaciju. Shematski prikaz strukture wwPDB-a dan je na slici 2.

Slika 2. Struktura baze podataka wwPDB (preuzeto s mrežne stranice baze wwPDB [50], 2017.).

Pored informacije o strukturama makromolekula baza PDB sadrži alate za pretraživanje proteinskih slijedova i njihovih eksperimentalno određenih struktura (određenih samostalno ili s

ligandima ili metalima). PDB sadrži i razne informacije u vezi mutacija, biotehnoloških modifikacija/preinaka lanaca u postupku izolacije ili kristalizacije, podatke o kvaliteti strukture, statističke podatke u vezi proteina i baze, te brojne poveznice na druge baze i bioinformatičke metode i servise za vizualizaciju i analizu proteinskih struktura. Nadalje, PDB omogućuje preuzimanje proteinskih slijedova u raznim oblicima (formatima), analizu sličnosti između primarnih struktura (slijedova), informacije o sekundarnoj strukturi, kao i analize sličnosti na temelju 3D strukture proteina. Iz baze PDB preuzete su u radu kompletne primarne strukture lanaca membranskih proteina radi analize sličnosti, razvoja i optimizacije algoritama za izbor reprezentativnih skupova membranskih proteina. Za daljnju obradu i provjeru preuzetih podataka i struktura, napisane su potrebne skripte i programi u programskom jeziku Python [47,48].

2.1.2. UniProt

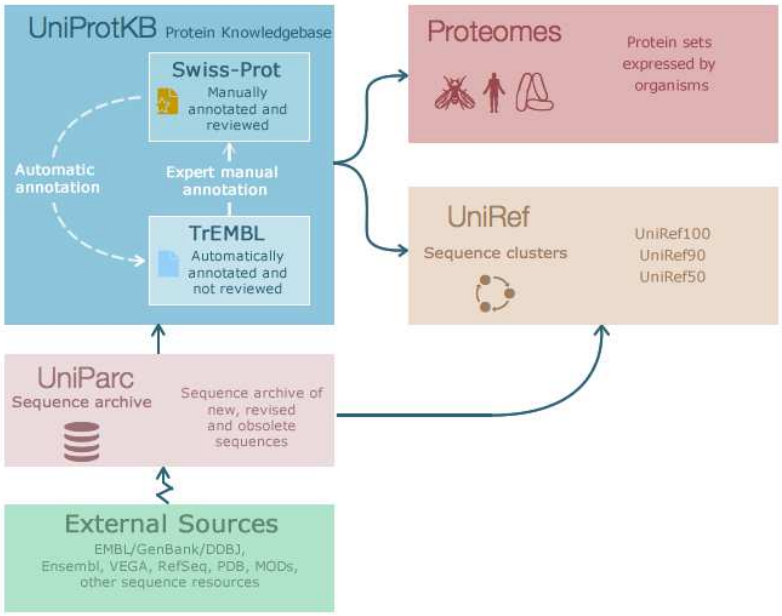
Konzorcij pod nazivom UniProt (engl. *The Universal Protein Resource*) [32] uspostavljen je 2002. godine a čine ga Europski institut za bioinformatiku (engl. *The European Bioinformatics Institute*, EMBL-EBI) [58], Švicarski institut za bioinformatiku (engl. *The Swiss Institute of Bioinformatics*, SIB) [59] i Baza proteinskih informacija (engl. *The Protein Information Resource*, PIR) [60]. Baza UniProt za glavni cilj ima pružanje, nadograđivanje i održavanje organiziranih informacija važnih za znanstvenu-stručnu zajednicu o slijedovima aminokiselina, strukturama, funkcijama, literaturnim izvorima i raznim bilješkama o proteinima. Također, UniProt sadrži poveznice na druge baze i programske metode za analizu i vizualizaciju proteina. Ti podaci smješteni su u nekoliko baza podataka poglavito u UniProtKB (engl. *UniProtKnowledge Base*) kao i UniRef (engl. *UniProt Reference Clusters*) [61], te UniParc (engl. *UniProt Archive*) [51].

UniProtKB (engl. *The UniProt KnowledgeBase*), središnja je baza (internetski servis) i repozitorij brojnih sustavno organiziranih informacija poput naziva, aminokiselinskog slijeda, podataka o funkciji, biokemijskim i metaboličkim mehanizmima u kojima protein sudjeluje, ligandima ili metalima s kojima protein gradi komplekse. Nadalje, UniProtKB sadrži literaturne izvore s poveznicama na druge baze koje sadrže genske informacije o proteinima, i na brojne programske alate poput onih za vizualizaciju ili modeliranje strukture i funkcije. Baza UniProtKB sastoji se od dva dijela: UniProtKB/Swiss-Prot [30,31] i UniProtKB/TrEMBL [32]. Swiss-Prot proteinska baza sadrži provjerene podatke prikupljene iz literature (preko pola milijuna proteinskih slijedova). Podaci su visoke kvalitete jer su u postupku prikupljanja i provjere pozorno provjeravani, te su u unosu navedene i sve važne dodatne napomene poput podataka o eksperimentalnoj metodi, uvjetima izvođenja eksperimenta, i slično. Visoki zahtjevi na točnost postavljeni u unosu novih zapisa u bazi SWISS-PROT ograničavali su brzinu unosa novih podataka. Stoga se pristupilo računalnom unosu novih proteinskih slijedova prevođenjem zapisa odgovarajućih kodirajućih genskih (nukleotidnih) slijedova iz baze TrEMBL koja sadrži preko 90 milijuna nukleotidnih slijedova.

UniRef [61], druga je (pod)baza unutar baze UniProt, a na internetskoj stranici daje proteinske slijedove iz baze UniProtKB i odabrane UniParc slijedove uz isključenje (skrivanje) onih koji imaju visoku međusobnu zalihost. UniRef pruža grupirane skupove slijedova radi dobivanja potpune pokrivenosti slijednog prostora na tri razine identičnosti i to:

- UniRef100 kombinira identične slijedove i pod-fragmente s 11 ili više aminokiselinskih ostataka iz bilo kojeg organizma u jedan unos UniRef (tj. klaster), s ciljem olakšavanja istraživanja bioloških osobina.
- UniRef90 je izgrađen grupiranjem (klasteriranjem) UniRef100 slijedova tako da se svaka grupa (klaster) sastoji od slijedova koje imaju najmanje 90% identičnosti slijeda, a 80% se preklapaju s najdužim slijedom u grupi,

- UniRef50 je izgrađen grupiranjem slijedova iz UniRef90 koji imaju najmanje 50% identičnosti slijeda, a 80% se preklapaju s najdužim slijedom u grupi.



Slika 3. Shematski prikaz strukture baze UniProt (preuzeto s mrežne stranice baze [32], 2017.).

UniParc [51] sveobuhvatna je baza podataka koja mapira i pohranjuje većinu javno dostupnih slijedova proteina u svijetu, uključujući i zastarjele podatke iz UniProtKB. Inače, identični proteini mogu biti pohranjeni u različite baze podataka po različitim kodnim (skraćenim) nazivima, kao i u više kopija u istoj bazi podataka. UniParc izbjegava takvu zalihost pohranjivanjem svakog jedinstvenog slijeda samo jednom. Pritom, zapisu pridjeljuje stabilan i jedinstveni identifikator (UPI) koji se ne mijenja niti uklanja. UniParc sadrži samo proteinske slijedove, a sve ostale informacije o proteinu moraju se preuzeti iz izvorne baze podataka. Podaci se osvježavaju svaka dva tjedna.

UniProt sadrži i informacije o molekularnoj funkciji, biološkim procesima, taksonomiji, položaju u stanici, zatim o izmjenama u aminokiselinama proteinskog slijeda, sekundarnoj strukturi, sličnosti između slijedova te o eksperimentalnim podacima i neusklađenostima u zapisima. Dodatno, postoji mogućnost traženja dijelova lokalne sličnosti pomoću programa BLAST, sravnjenja proteinskih slijedova kao i mogućnost odabira različitih formata u prikazu ili preuzimanju informacija (tekst, fasta, xml, gff, rdf i sl.).

2.1.3. Baza proteina smještenih u membranu OPM

OPM baza proteina sadrži izdvojeni podskup proteina koji su u interakciji s membranom, bilo da interagiraju s površinom membrane ili s njenim unutarnjim lipidnim dijelom [28,29], a strukture su preuzete iz baze PDB [20].

Podaci o proteinima u bazi OPM razvrstani su na tri glavne podskupine, i to na strukture:

a) *transmembranskih (TM) proteina*, a trenutačno ih ima oko 1700,
b) *perifernih membranskih proteina* (oko 1340 proteina, 1570 podjedinica, i nemaju TM segmente nego su pričvršćeni na površinu membrane), te
c) *peptida* (516 proteina) koji imaju 609 podjedinica (imaju ukupno 83 segmenta pravilne strukture koji nisu pravi TM segmenti nego su na različite načine položeni u membranu).

U radu na disertaciji rabljeni su strukturni podaci za skupinu TM proteina. Među 1706 proteina s oko 5400 podjedinica (lanaca, slijedova), njih 1414 je s jednim ili više TM segmenata. U tih 1414 proteina ima ukupno oko 22370 TM segmenata (α ili β vrste).

Podskupina TM proteina u bazi OPM dalje se dijeli na tri pod-skupine:

a) 1096 proteina (s ukupno 4180 lanaca/podjedinica, i 17479 TM segmenata) koji sadrže više od jednog TM segmenta alfa sekundarne strukture (TM_polytopic-alpha, engl. *alpha-helical polytopic class*), najveća je podskupina membranskih proteina, koja je ujedno najviše analizirana u razvoju metoda za predviđanje strukture membranskih proteina),
b) 387 proteina (s ukupno 758 lanaca) čiji svaki lanac sadrži jedan TM segment alfa vrste (TM_bitopic-alpha, engl. *alpha-helical bitopic class*),
c) 224 kraća proteina (peptida) s ukupno 486 lanaca i 4464 TM segmenata koji poprimaju beta sekundarnu strukturu.

Pregledom pojedinačnih proteina u tim podskupinama, uočena su neka odstupanja. U podskupini TM_polytopic-alpha pregledom 1096 TM proteina s lancima koji imaju više od jednog TM segmenta alfa vrste uočeno je kako za 31 protein (s ukupno 312 lanaca) nije naveden niti jedan TM segment u bazi OPM. Dodatno za 20 TM proteina ima naveden samo po jedan TM segment, pa bi na temelju toga trebali biti u drugoj skupini TM_bitopic-alpha. Kad se sve to uzme u obzir, najveća skupina TM proteina (TM_polytopic-alpha) s dva ili više TM segmenta alfa vrste zapravo ima 1045 TM proteina.

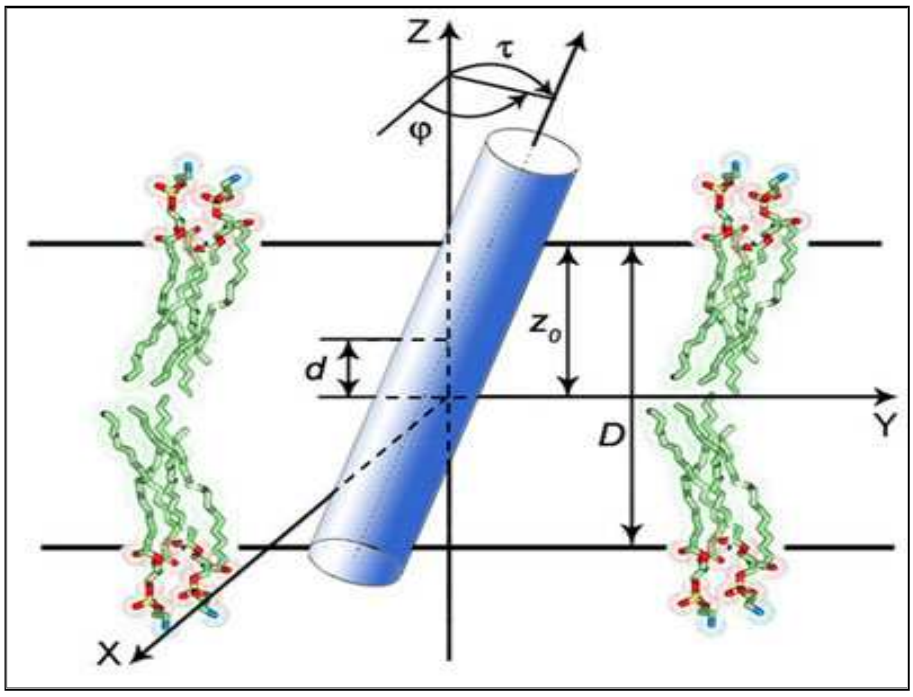
U nekih proteina načinjene su jednostavno uočljive pogreške u unosu u bazu OPM, poput strukture transportnog proteina CmeB PDB koda 5lq3, kod kojega je zamijenjen broj podjedinica i broj TM segmenata u zbirnoj tablici. Naime, navedeno je da protein 5lq3 ima 36 podjedinica s (ukupno) 3 TM segmenta alfa vrste, a zapravo je obrnuto, tj. protein ima 3 podjedinice (lanca) s ukupno 36 TM segmenata. Ista pogreška u zbirnoj tablici načinjena je kod proteina 5weo (koji stvarno ima 4 lanca svaki po 7 TM segmenata i ukupno 28 TM segmenata) i 3jc2 (koji stvarno ima 4 podjedinice i ukupno 13 TM segmenata). Nadalje, kod strukturno skoro identičnih proteina (različiti zapisi citokromskih kompleksa) 3cxh, 1ezv, 1bcc, 3bcc u zbirnoj je tablici navedeno da imaju ukupno 24 lanca s 10 TM segmenata, a kad se pogledaju same strukture vidi se kako su zapravo ti brojevi zamijenjeni. Kod sva ta četiri proteinska kompleksa (3cxh, 1ezv, 1bcc, 3bcc) strukture se sastoje od 8 lanaca od kojih svaki ima pojedan TM segment i od dva lanca od kojih svaki ima po 8 TM segmenata. Naposljetku, takva pogreška načinjena je u bazi OPM i kod proteina 5xtc koji ima 20 podjedinica s (ukupno) 80 TM segmenata.

Nakon ovih ispravki, najvažniji skup TM proteina (korišten kao početni skup u disertaciji), u kojem svaki proteinski lanac ima više odjednog TM segmenta alfa sekundarne strukture (skup naziva TM_polytopic-alpha), trebao bi sadržavati 1045 proteina s 3537 podjedinica i 17510 TM segmenata. Pored skupa TM_polytopic-alpha u analizama u disertaciji za sastavljanje početnoga skupa membranskih proteina alfa vrste koristila se i podskupina TM_bitopic-alpha. Kod te skupine treba napomenuti kako 248 od 387 proteina (s 335 podjedinica) nema u OPMu navedene TM segmente, pa se takvi proteini trebaju svrstati u skupinu površinskih proteina. Nakon ispravke broja proteina i broja podjedinica, u skupini proteina TM_bitopic-alpha nalazi se 139 proteina s 423 lanca i 430 TM segmenata alfa vrste. Kad tu podskupinu spojimo sa skupinom TM_polytopic-alpha, za izradu početnog skupa dobije se ukupno *1204 proteina* s *4110 lanaca* (podjedinica) koji imaju ukupno *18090 TM segmenata alfa vrste*. Međutim, među tim podjedinicama nalazi se veliki broj identičnih ili skoro identičnih aminokiselinskih slijedova od kojih se za kasnije analize zadrži samo jedan lanac.

S ciljem analize, preuzeti su dokumenti s mrežne stranice baze OPM [28,29] s položajima TM segmenata u proteinskim slijedovima, kao i liste proteina, te su isti pridruženi slijedovima i eksperimentalnim strukturama preuzetim iz baze PDB. Pritom, reprezentativni je protein onaj protein koji predstavlja neku skupinu proteina (superobitelj, obitelj ili skup istih proteina), i to nije protein koji će se u konačnici nužno izabrati u reprezentativni skup. U ovom dijelu oni se

odabiru kako bi se smanjila zalihost početnog skupa proteina iz baze PDB. Reprezentativni protein bira se iz skupa proteina tako što se gleda koji od njh ima najcjelovitiju eksperimentalnu strukturu s najvećim brojem proteinskih domena i manje dijelova s neuređenom strukturom. Ako struktura istog proteina odgovara različitim konformacijskim stanjima ili kvarternim kompleksima proteina, tada i oni ulaze u izbor kao reprezentativni proteini.

Ako se prilikom izbora proteina niske međusobne sličnosti javljaju proteini identičnih slijedova, tada se na osnovu podatka o eksperimentalnoj metodi kojom je struktura određena i bolje rezolucije donosi odluka koji se protein zadržava u početnom skupu.



Slika 4. Shematski prikaz transmembranskog proteina u hidrofobnoj ploči (preuzeto sa stranica baze OPM [29], 2017.).

Glavna korisna informacija koju donosi baza OPM u odnosu na strukturu proteina koja se može preuzeti iz baze PDB je optimizacija smještanja membranskog proteina u membranu. Smještanje proteina u lipidni dvosloj promjenjive debljine provodi se minimizacijom slobodne energije prijenosa proteina pri prijelazu iz vode u membranu $\Delta G_{transfer}$ [52] (anizotropni solvatacijski model lipidnog dvosloja [53]). Minimizacija se provodi, u ovisnosti o varijablama $d,\ z_0, \varphi$ i τ (slika 4) definiranim u koordinatnom sustavu čija se z os podudara s okomicom na lipidni dvosloj.

Na slici 4. danje shematski prikaz transmembranskog proteina u hidrofobnoj ploči (lipidnom dvosloju), gdje je d pomak duž okomice na lipidni dvosloj koji se optimira; D je debljina hidrofobnog sloja ($D = 2z_0$); φ je kut rotacije, a τ je kut nagiba proteina ili njegovog dijela u odnosu na os z.

2.1.4. Baza transmembranskih proteina PDBTM

PDBTM (engl. *Protein Data Bank of Transmembrane Proteins*) druga je specijalizirana baza za TM proteine izrađena i održavana od mađarske istraživačke grupe s Instituta za enzimologiju [54,55,56]. Baza podataka PDBTM [56] stvorena je pregledom i izdvajanjem svih proteinskih struktura iz PDB s pomoću algoritma TMDET [57]. Taj algoritam smješta aminokiselinski slijed

u odnosu na membranu, slično kao što algoritam DSSP (engl. *Definition of Secondary Structure of Proteins*) definira sekundarnu strukturu proteina [62].

Algoritam TMDET koristi datoteke proteinskih struktura iz baze PDB kao ulazne podatke. Prvo se ispituje vrsta proteina i lanaca (pritom se isključuju proteini iz virusa) i zanemaruje proteinske slijedove kraće od 15 aminokiselinskih ostataka. Strukture proteina koje imaju samo koordinate C_α atoma gledaju se odvojeno, jednako kao i strukture određene s nedovoljno dobrom rezolucijom. Osnova je algoritma traženje najvjerojatnijeg položaja membranskih ravnina u odnosu na zadane koordinate atoma. Radi usporedbe i provjere usuglašenosti baza, u disertacijije korišten skup od 190 lanaca iz [42] s oznakama sekundarne strukture alfa vrste prema bazi OPM kao i s oznakama prema bazi PDBTM [56].

2.2. Program Emboss s aplikacijom *needleall* za analizu sličnosti među proteinima

EMBOSS [49] (engl. *The European Molecular Biology Open Software Suite*) besplatan je programski paket otvorenoga koda za analizu slijedova koji je posebno razvijen za potrebe zajednice istraživača iz područja molekularne biologije.

Programski paket instaliran je na računalo i rabljen u analizama. Korištena je aplikacija pod nazivom "needleall" za globalno sravnjivanje svih parova primarnih struktura proteina iz zadanoga skupa od *N* proteina. EMBOSS se temelji na Needleman-Wunsch-ovom algoritmu [63] za sravnjenje proteinskih slijedova. Program ima više mogućnosti korisničkih odabira poput matrica sličnosti (zamjene) aminokiselina, otvaranje ili produženja prekida (engl. *gap*) u proteinskom slijedu radi ukupne optimizacije sravnjenja, različitih izlaznih formata, itd. Pri sravnjivanju slijedova u disertaciji koristila se matrica EBLOSUM 62 te uobičajene (ponuđene) postavke za sravnjenje, ili postavke koje su koristili drugi istraživači poput grupe E. Rath [41].

2.3. Algoritmi za smanjenje zalihosti skupova razvijeni od strane drugih autora

Pregledom literature pronađena su tri novija znanstvena rada u kojima su provođeni postupci izbora reprezentativnih skupova membranskih proteina niske zalihosti (niže od 30%) [41,42,43]. U tim su radovima rabljeni algoritmi Hobohm 2 [24] i UniqueProt [44]. Osim ova dva algoritma u literaturi se navodi samo još jedan algoritam nazvan Hobohm 1 [24]. To je ujedno i prvi algoritam razvijen za potrebe izbora skupova proteina niske zalihosti u primarnoj strukturi, a razvijen je u vrijeme kad je u cijeloj bazi PDB postojalo manje od 1000 struktura.

2.3.1. Algoritmi Hobohm 1 i Hobohm 2 za smanjenje zalihosti među proteinima

Dva algoritma najprije nazvani algoritam 1 i algoritam 2 (kasnije prozvani Hobohm 1 i Hobohm 2) za dobivanje skupova proteinskih lanaca čija je međusobna sličnost niža od unaprijed definiranoga praga sličnosti (ili identičnosti), opisani su u ref. [24].

Algoritam Hobohm 1 radi (po načelu 'odabiri dok radi') na sljedeći način:

1. Odabire se proizvoljni protein iz početne, po nekom kriteriju uređene (sortirane), liste proteina i odbacuju se u tom koraku svi proteini koji su odabranom proteinu slični više od unaprijed zadnoga (najvišeg dopuštenog) praga sličnosti (zadnoga od strane korisnika).
2. Zatim algoritam odabire sljedeći protein i nastavlja na analogan način sve dok se uređena lista proteina ne iscrpi. Postupak završava kad preostanu samo oni proteini koji nemaju u skupu niti jedan protein koji bi im bio sličan više od unaprijed izabranog najvišeg praga sličnosti.

3. Odbacuju se svi oni proteini koji ne zadovoljavaju standarde odabrane od strane korisnika (npr. rezolucija – kvaliteta podataka).

Algoritam Hobohm 2 radi (po načelu 'uklanjaj dok radi') na sljedeći način:

1. U zadanom popisu (listi, skupu) proteinskih kandidata (poredanih/sortiranih po unaprijed odabranome kriteriju) za svaki protein traži se ukupni broj proteina (broj susjeda) koji su odabranome proteinu slični više od zadanoga praga sličnosti. Nakon toga, uklanja se jedan protein iz skupa i iz daljnjih razmatranja (i uklanja se s popisa susjeda za sve preostale proteine).
2. Potom se postupak nastavlja sve dok u preostalom skupu ne bude više proteina koji imaju susjednih (koji bi mu bili slični više nego je definirani najviši prag sličnosti). S obzirom da je broj veza (susjednih poveznica) za početni skup stalan, s ciljem povećanja broja preostalih susjeda u konačnom reprezentativnom skupu, mogu se za izbacivanje odabirati upravo oni proteini koji imaju najveći broj veza. Na taj način iscrpit će se sve veze uz izbacivanje najmanjeg broja proteina, tj. preostat će najveći broj proteina koji su međusobno nezalihni. Svakako da korisnik prije pokretanja algoritma može odlučiti hoće li neke od proteina uključiti u odabir ili isključiti iz odabira. Određeni proteinski lanci za koje se smatra da su od izrazite važnosti mogu se obilježiti za ostavljanje u reprezentativnom skupu.

U istom radu navedeno je kako je algoritam 2 (Hobohm 2) superiorniji ukoliko je cilj izbor najvećeg broja proteinskih lanaca u konačnom reprezentativnom skupu.

Nakon što su opisani i definirani algoritmi Hobohm 1 i 2 [24], u sljedećem radu [25] opisana je njihova nadogradnja s novim kriterijima za odabir proteinskih lanaca koji su obilježeni za ostaviti za kasnija razmatranja, odnosno za isključiti iz razmatranja i iz reprezentativnog skupa. Kriteriji po kojima se radi obilježavanje su:

(a) prvo se izbacuju proteini koji su obilježeni za isključiti,
(b) zatim se izbacuju oni lanci koji imaju lošiju kvalitetu,
(c) ako je više onih s najnižom (istom) kvalitetom izbacuje se onaj lanac koji nije označen za ostaviti,
(d) ako je izbor i dalje višestruk, izbacuje se onaj s višim PDB kodom (gledano abecedno ili brojčano, pa tako npr. ako imamo proteine s PDB kodovima 1M01 i 2PRC, izbacio bi se 2PRC jer u uzlaznom abecednom ili brojčanom slaganju/redanju 2PRC slijedi nakon 1M01),
(e) konačno, ako je izbor još uvijek višestruk, izbacuje se onaj koji ima višu abecednu ili brojčanu oznaku lanca (npr. ostavio bi se 3WU2_A u odnosu na 3WU2_B).

U drugom kriteriju (b) navedeno je da se izbacuju oni proteinski lanci koji imaju lošiju kvalitetu strukture, a kvaliteta je definirana parametrom Q danim sljedećim izrazom:

$$Q = r + \frac{R}{20}$$

gdje je s r – označena rezolucija izražena u angstremima (Å), R je R-faktor, koji je mjera slaganja između kristalografskog modela i eksperimentalnih podataka dobivenih difrakcijom rendgenskih zraka, i izražava se u postotcima. Proteinski lanac s većim Q zapravo je lanac koji ima strukturu niže kvalitete, i takvi se lanci nastoje odbaciti prilikom rada algoritama za izbor reprezentativnog skupa membranskih proteina.

2.3.2. Algoritam UniqueProt za smanjenje zalihosti među proteinima

Program UniqueProt razvili su S. Mika i B. Rost [44] za izbor reprezentativnog skupa proteina, točnije proteinskih lanaca/podjedinica. Sama kratica (naziv) programa upućuje na to da se tim algoritmom nastoje izabrati jedinstveni (engl. *Unique*) proteini. Kako bi se numerički iskazala sličnost među proteinskih slijedova, algoritam računa i rabi u radu HSSP (engl. *Homology-derived Secondary Structure of Proteins*) vrijednosti za svaki lanac [21]. Vrijednost HSSP (engl. HSSP-*value*, što je skraćeno kao *HV*) računa se iz postotne identičnosti (PID) dvaju lanaca, koja se dobije kao postotak broja identičnih ostataka u sravnjenju proteinskih lanaca programom BLAST [22,23], i duljine njihovog sravnjenja L, prema sljedećem izrazu:

$$\text{HSSP-vrijednost(L;PID)} = PID - \begin{cases} 100 & L \leq 11 \\ 480 \cdot L^{-0.32 \cdot \{1+exp(-L/100)\}} & L \leq 450 \\ 19.5 & L > 450 \end{cases}$$

Pritom se u izračunu duljine L ne uzimaju u obzir praznine (engl. *gap*) otvorene u proteinskim lancima. Praznine se otvaraju prilikom sravnjenja slijedova s ciljem optimizacije analize i kvantificiranja sličnosti između slijedova kako bi se uzele u obzir i sličnosti u manjim segmentima, tj. dijelovima lanaca. Kako vidimo, kod ovog načina izračuna sličnosti uzima se u obzir duljina sravnjenja radi toga što je (u grubo) veća nasumična vjerojatnost da se pojavi sličnost kod kraćih nego kod dužih proteinskih lanaca (ili segmenata). Odnosno, ukoliko u jednom od dva proteinska lanca za koja želimo analizirati sličnost (za koje je duljina sravnjenja L) izdvojimo segment (isječak) duljine l aminokiselina, onda je vjerojatnost da ćemo isti takav segment pronaći u lancu drugog proteina to veća što je taj segment kraći (tj. l što manji) i što je taj drugi lanac dulji, (tj. L veći). U gornjem izrazu konstante u eksponentu u srednjem području duljina sravnjenja (za $L \leq 450$) kao i konstante za niske i visoke L dobivene su empirijski, tj. ugađanjem krivulje ovisnosti postotne sličnosti PID o duljini sravnjenja L. Ta krivulja razgraničava područje visoke identičnosti među lancima $HV > 0$ (u kojima se javlja i sličnost u strukturama među lancima) i područje niske identičnosti ($HV < 0$) gdje nema sličnosti u strukturama među uspoređivanim lancima [64]. Nadalje, gornja krivulja uvažava opažanje temeljeno na analizi poznatih struktura da za kratke lance (duljine sravnjenja $L < 11$) i jako visoka identičnost (visok PID) ne ukazuje na sličnost u strukturama. Za $HV = 0$ dobivamo krivulju ovisnosti identičnosti PID o duljini sravnjenja između proteinskih lanaca L poput one iz 1991. kada su Sander i Schneider [21] ustanovili da za sravnjenja duljine veće od 80 ($L > 80$) krivulja PID *vs* L dolazi u zasićenje za vrijednost PID = 25%. Za dva proteinska lanca za koja se u analizi sličnosti dobije $HV > 0$, kaže se da su bliski (slični), a u slučaju ako je $HV < 0$ kaže se da su udaljeni. Taj je kriterij rabljen u algoritmu UniqueProt [44] za odlučivanje u svakom koraku rada algoritma za izbor reprezentativnog skupa proteina niske sličnosti (manje od praga sličnosti).

Za razliku od algoritma UniqueProt koji u radu koristi $HV = 0$, tj. krivulju PID *vs* L kao razgraničenje između parova proteinskih lanaca koji su slični i onih koji nisu slični, algoritmi Hobohm 1 i 2 u svome radu kao granicu rabe konstantu iznosa jednakog definiranom pragu identičnosti/sličnosti. Te su vrijednosti obično u rasponu 20 – 30% za sve duljine sravnjenja L. Tako će i za kratke lance (npr. $30 < L < 50$) granica odbacivanja biti u rasponu identičnosti/sličnosti 20 – 30%, što će dovoditi do njihovog odbacivanja iz konačnog skupa proteinskih lanaca. U nastavku ilustriramo razliku u kriterijima između algoritama UniqueProt [44] i Hobohm 2 [24] na sljedećem primjeru:

a) postavimo prag identičnosti od 20% prema algoritmu Hobohm 2,
b) izaberimo za analizu dva kraća lanca za koje je duljina sravnjenja $L = 50$,
c) neka je utvrđena postotna identičnost PID $= 40\%$,

d) prema UniqueProt dobivamo $HV = -1.73$ i s obzirom da je $HV < 0$ kod takvoga para lanaca ne postoji sličnost u strukturama i niti jedan lanac neće biti izbačen radi njihove međusobne sličnosti/identičnosti,
e) međutim, prema algoritmu Hobohm 2 jedan od ta dva proteinska lanca mora biti izuzet iz konačnog reprezentativnog skupa na toj razini međusobne usporedbe, jer je između lanaca postotna identičnost PID = 40% (> 20%, što je prag identičnosti).

Nadalje, provjerimo kakva je situacija kod duljih slijedova u rasponu duljina sravnjenja (L) 300 do 450 aminokiselinskih ostataka, za prag identičnosti od 20% i 19.5%. Za duljinu sravnjenja 300, i jedna i druga metoda daju identičnosti koje ne prelaze prag od 20%. Kod duljina sravnjenja u rasponu 310 do 450 metoda UniqueProt malo je strožija i odbacuje slijedove čija je postotna identičnost 20%. Međutim, ukoliko je identičnost 19.4% za sve duljine sravnjenja veće od 310 i metode Hobohm 2 i UniqueProt pokazuju jednake strogosti. Ove usporedbe potvrđuju kako je kod kratkih slijedova kriterij stalnog praga identičnosti strožiji nego kriterij HV (tj. $HV < 0$, kada nema identičnosti među lancima) u slučaju algoritma UniqueProt. Metode koje koriste stalni prag identičnosti poput metode Hobohm 2 samo su malo manje stroge (za svega 0.6%) kod duljih sravnjenja, tj. onih većih od 310 aminokiselinskih ostataka. Napomenimo da kod usporedbi dvaju duljih proteinskih slijedova, duljina sravnjenja bez uzimanja u obzir praznina (u pravilu) odgovara duljini slijedova.

U nastavku bit će ukratko opisani osnovni koraci i logika odlučivanja u radu algoritma UniqueProt u postupku izbora reprezentativnog skupa proteinskih lanaca. Nastoji se, u konačnici, pronaći najveći podskup (početnog skupa) lanaca koji zadovoljava uvjet da niti jedan par u tom podskupu nema $HV > 0$, gdje je u tom algoritmu $HV = 0$ postavljeno kao granica odlučivanja. Potom se za svaki protein P u početnom skupu izbroji broj proteina NP koji s proteinom P imaju $HV > 0$. Smatra se da svi proteini {NP} s $HV > 0$ pripadaju obitelji F(P), koji se još nazivaju i susjedima proteina P. Zatim se za svaki protein pohranjuje broj i kodovi svih susjeda i cijeli skup poreda (sortira) prema veličini obitelji {F}. Konačno, uzmimo da algoritam kreće od proteina P' koji ima najbrojniju (najveću) obitelj {F}, on u tom koraku isključuje sve članove obitelji F(P'). U nastavku su opisane situacije na koje algoritam može naići tijekom odlučivanja:

1) ako obitelj F(P') sadrži samo jedan slijed (lanac), dodaje se P' u popis jedinstvenih proteina (a time i u konačni skup proteina među kojima nema zalihosti u odnosu na unaprijed definirani kriterij, tj. prag identičnosti),
2) ako {F (P')} ima više od jednog proteinskog lanca, brišu se s popisa svi članovi obitelji osim P',
3) ako P' ima u obitelji jednog člana Q koji je već uključen u jedinstveni popis u prethodnom koraku, predstavnik P' i svi drugi članovi obitelji {F(P')} osim Q bit će uklonjeni iz preostalog popisa kandidata. Ova situacija može imati dva razloga:
   (a) zbog asimetričnosti matrice udaljenosti (koja kao elemente za svaki par sadrži vrijednosti HV) i
   (b) radi nekih preklapanja između domena koje poništavaju trokutni odnos sličnosti među proteinima A, B i C (tj. iz A sličan B i A sličan C ne znači da je B sličan C). Algoritam završava kada nijedan lanac ne ostane u preostalom popisu (tj. radnom skupu koji se analizira u svakom koraku izvršenja algoritma). Drugim riječima, algoritam završava kad se iscrpe proteini u radnom skupu ili ako se dogodi da u radnom skupu svi parovi proteina imaju identičnost ispod praga (tj. $HV < 0$),
4) kad se u radu algoritma odlučuje između dva proteina s $HV > 0$, prednost u zadržavanju daje se duljem lancu.

Algoritam se može krenuti izvršavati i po obrnuto poredanoj listi, tj. kad se na prvo mjesto u uređenoj listi stavi protein čija obitelj ima najmanje sličnih lanaca (članova, susjeda).

2.4. Skupovi za usporedbu s drugim metodama

U ovom odjeljku opisat će se polazni skupovi podataka izdvojeni iz baza proteinskih slijedova i struktura, kao i reprezentativni skupovi dobiveni prethodno prikazanim algoritmima od strane drugih autora [41,42,43]. Ti skupovi korišteni su u razvoju algoritama u disertaciji, te u njihovim poboljšanjima i usporedbama s rezultatima i algoritmima iz literature.

2.4.1. Opis skupa M190

Istraživačka grupa profesora Burkharda Rosta provela je izbor reprezentativnog skupa membranskih proteina analizom tada raspoloživih primarnih struktura (lanaca, aminokiselinskih slijedova) membranskih proteina u bazama PDB [20] i UniProt [32] uz prag identičnosti 20%. Sekundarna struktura, položaji TM segmenata i topologija membranskih proteina preuzeti su iz dvije baze specijalizirane za membranske proteine - OPM [28,29] i PDBTM [56]. Primarne strukture izdvojene su iz 462 TM proteina (TMP) koji su se nalazili u tim bazama u rujnu 2013. godine. Ti su proteini sadržavali ukupno 1101 proteinski lanac (podjedinicu) koji imaju cjelovite slijedove i jedinstvenu povezanost između baza PDB [20] i UniProt [32].

Identičnost među podjedinicama i izbor reprezentativnog skupa provedena je u radu [42] uporabom ranije razvijenoga algoritma UniqueProt [44], uz uvjet da je vrijednost $HVAL > 0$. U konačnici dobiveno je 190 proteinskih lanaca u skupu s identičnošću manjom od 20%. U sastavljanju početnoga skupa birani su proteinski lanci u kojima je manje neodređenih aminokiselinskih ostataka, kao i dulji lanci.

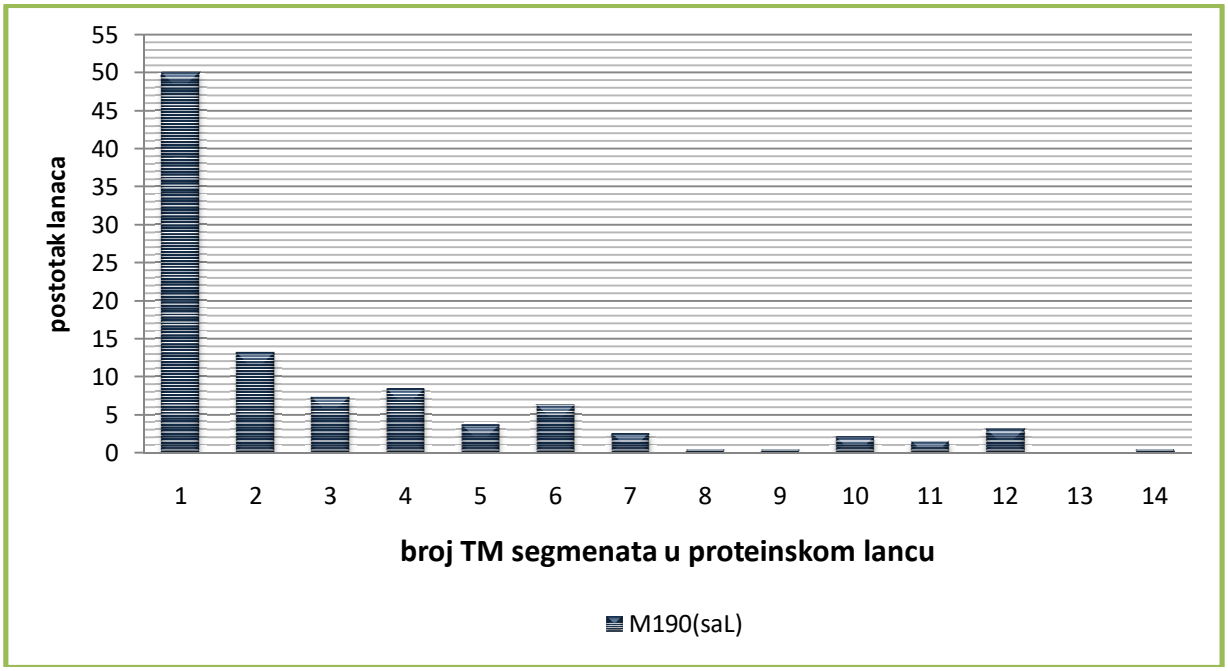
Ovim postupkom dobiven je skup bez zalihosti u kojima je većina struktura proteina određena eksperimentalno metodom rendgenske difrakcije X-zraka (rezolucija do 9 Å). Prosječna rezolucija u skupu iznosi 2.9 Å. Skup sadrži 569 TM segmenata (prosječno ~ 3 po proteinskom lancu) s 50179 aminokiselinskih ostataka, a prosječna je duljina proteinskoga lanca 264 aminokiseline. Karakteristični podaci skupa M190 dobivenog uz prag identičnosti 20% [42] prikazani su u tablici 2.

Tablica 2. Osnovne osobine skupa M190 iskazane po lancu i za cijeli skup.

skup M190	AK	TM segment	AK u membrani	% AK u membrani
minimum po lancu	29	1	6	1.67
maksimum po lancu	1342	14	305	78.13
srednja vrijednost po lancu	264.10	2.99	63.83	35.19
ukupno – skup	50179	569	12127	

AK – aminokiselina, TM – transmembranski

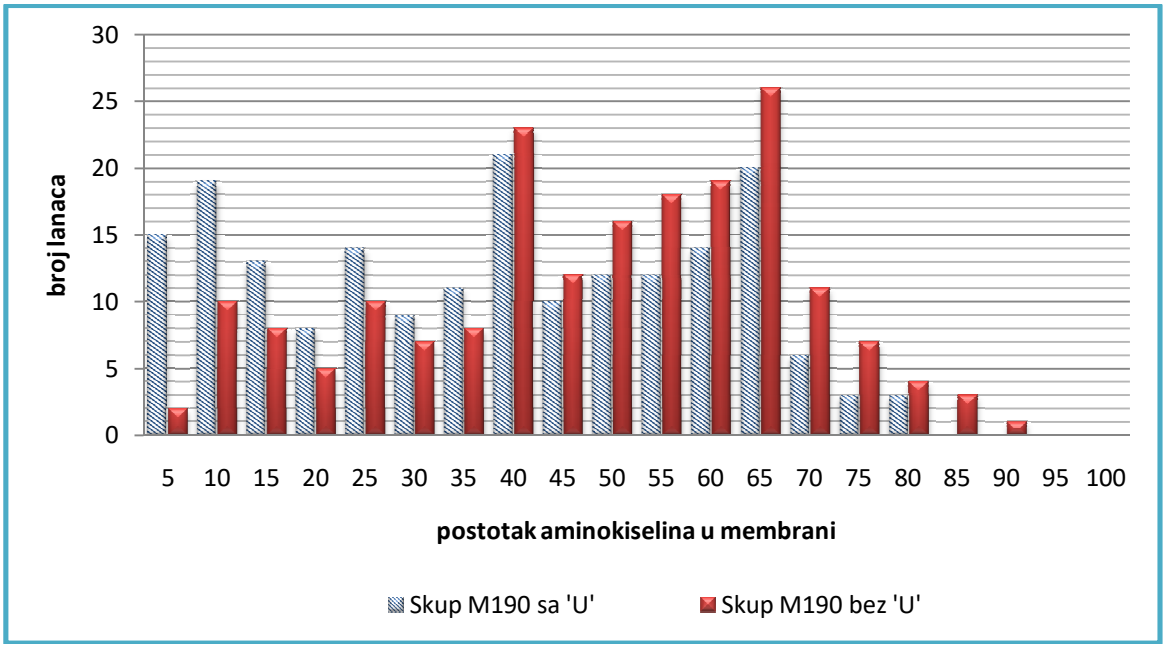
U postupku izbora ovoga skupa razmatrani su samo oni dijelovi slijedova koji se u potpunosti slažu u bazama UniProt i PDB. Kako su proteinski slijedovi u bazi UniProt [32] u pravilu cjelovitiji i dulji nego oni u bazi PDB [20], a struktura je riješena samo za lanac koji je naveden u bazi PDB, za taj dodatni dio lanca preuzet iz baze UniProt sekundarna struktura je označena kao nedefinirana (engl. *undefined* ili skraćeno 'U'). U analizama i izračunima taj se dodatni dio slijeda ne uzima u obzir, a kao TM segmenti uzimaju se i oni koji su zapravo izvan-membranske petlje, i označavaju se s 'L' (engl. *loop*). Na slici 5 prikazana je raspodjela broja TM segmenata po lancima u skupu M190 (koji uključuju i lance u kojima ima dijelova sekundarne strukture 'L').



Slika 5. Postotak proteinskih lanaca u početnom skupu M190 po podskupovima lanaca istog broja TM segmenata.

Vidi se kako ovaj reprezentativni skup proteina ima 95 lanaca (50% od ukupnog broja) s jednim TM segmentom. Ta činjenica vodi na pitanje: je li tako veliki broj lanaca s jednim TM segmentom posljedica odabira algoritma koji nastoji dobiti maksimalni broj lanaca u reprezentatvnom skupu, ili je to posljedica načina analize i kvantificiranja same sličnosti među proteinskim lancima? Odgovor će se pokušati dobiti usporedbom i analizom različitih algoritama na više konačnih reprezentativnih skupova. Zasad je jasno kako je jedan dio razloga za tako velik broj proteina s jednim TM segmentom (koji su u pravilu i kraći lanci) i u tome što algoritam UniqueProt dopušta višu postotnu identičnost kod kraćih slijedova. Naime, kod kraćih slijedova *HV* biva manji od nule i kod postotne identičnosti (PID) veće od 20%. Tako na primjer, kod slijeda (duljine sravnjenja bez praznina) od 50 aminokiselinskih ostataka PID može biti i 40% (odnosno, za slučaj slijeda duljine 80 PID može biti do 32%, a kod slijeda duljine 100 PID može biti do 28%) a da u svakom od tih slučajeva *HV* bude manji od nule. Takvi se slijedovi (lanci) po metodi UniqueProt [44] smatraju identičnima manje od definiranoga praga (*HV* = 0, odnosno PID ~ 20%).

Problematika razmatranja lanaca u kojima dolazi do preklapanja u zapisima iz baza PDB i UniProt najbolje se vidi na primjeru lanca 1a11_A koji ima 517 aminokiselina i jedan TM segment duljine 18 aminokiselina. Ako se uzme dio slijeda bez 'U' (s 'U' se označavaju mjesta aminokiselina koja se ne poklapaju u zapisima u tim bazama) onda se slijedovi u zapisima podudaraju u 21 aminokiseline. Ukoliko se zanemare dijelovi lanaca koji se ne podudaraju, udio aminokiselina koje pripadaju TM segmentu u ukupnom lancu povećava se s 3.48% na 85.71%. Zbog ovih razlika, na slici 6 dana je raspodjela broja proteinskih lanaca po postotnom udjelu aminokiselina unutar TM segmenata, i to kada se uzmu u obzir cijeli proteinski slijedovi (sa U), i kada se uzimaju u obzir samo oni dijelovi slijedova koji se podudaraju u zapisima u obje baze (bez U). Vidi se da ako se u analizama uzimaju u obzir i dijelovi proteinskih slijedova koji se ne slažu između baza PDB i UniProt, broj lanaca s postotnim udjelom aminokiselina u membrani manjim od 40% značajno poraste (s 50 lanaca na 89 lanaca), odnosno s 26.32% na 46.84%.



Slika 6. Raspodjela broja proteinskih lanaca po postotnom udjelu aminokiselinskih ostataka u membrani za skup M190 (sa L).

Ako se promatra skup uz zanemarenje dijelova proteinskog slijeda koji se razlikuju u bazama PDB i UniProt, značajan je broj lanaca s postotnim udjelom aminokiselina između 40 i 75%. Prilikom analize sličnosti/identičnosti među proteinima iz skupa M190 koristit će se program EMBOSS s aplikacijom *needleall*. Nadalje, s obzirom da je skup M190 dobiven metodom UniqueProt, taj će skup biti korišten za usporedbu razina strogosti (pojašnjenje u nastavku) algoritama razvijenih u disertaciji i algoritma UniqueProt. Promotrimo dva algoritma (Algoritam 1 i Algoritam 2) koji analizu i kvantificiranje sličnosti među proteinima i izbor reprezentativnih skupova (RS) proteina provode različitim pristupima (polazeći od istoga početnoga skupa). Označimo s N_1 i N_2 broj izabranih proteina u konačnim reprezentativnim skupovima (nazovimo ih skraćeno RS1 i RS2). Ti se reprezentativni skupovi ne razlikuju samo u ukupnom broju proteina, nego se razlikuju i u samim vrstama proteina koji su izabrani u konačne skupove (tj. manji skup nije podskup većega). Ukoliko algoritam 1 primijenimo na RS2 i on pritom ne pronađe suvišnih (ne-jedinstvenih) lanaca, a algoritam 2 primijenjen na RS1 pronađe određeni broj suvišnih proteina, onda kažemo (u tom smislu) da je algoritam 2 strožiji algoritam nego algoritam 2. Radi toga, a kako bi se moglo provjeriti detaljnije razinu strogosti razvijenih algoritama i dviju realizacija algoritma UniqueProt iz literature [42,43], izdvojena su još tri skupa membranskih proteina (za različite razine sličnosti ili identičnosti). Polazeći od tih skupova proveden je izbor reprezentativnih skupova programom UniqueProt (ljubaznošću M. Bernhofera). Treba napomenuti da algoritam UniqueProt za različite ulazne parametre u programu (npr. redoslijed kojima su poredani proteini u ulaznom skupu i po kojem ih algoritam razmatra) daje značajno različite konačne rezultate (reprezentativne skupove proteina). Stoga su u svim provjerama i analizama tim algoritmom, ulazni parametri postavljeni na način kako je to opisano u najnovijem radu iz literature [43].

2.4.2. Opis skupa S481 (i njegovog podskupa S392)

U radu "grupe Sydney" [41] definiran je skup proteinskih lanaca izdvojenih iz baze OPM u kojima su u veljači 2012. godine (po autorima) bila prisutna 1045 jedinstvena proteinska slijeda. Od toga je 481 slijed iz strukturne podskupine *bitopic* (s jednim TM segmentom) ili *polytopic* (s dva ili više TM segmenata) s TM segmentima u sekundarnoj strukturi alfa uzvojnice. Nadalje, 95

proteinskih slijedova ima TM segmente beta sekundarne strukture, a preostalih 469 slijedova topljivi su proteini (koji su u interakciji s membranom, ali u nju nisu uronjeni). Pritom su uzeti u obzir lanci (integralnih ili površinskih) membranskih proteina čije su strukture određene (jednom od eksperimentalnih metoda):

- rendgendskom difrakcijom (engl. *X-ray*),
- nuklearnom megnetskom rezonancijom (NMR) u otopini (engl. *solution NMR*),
- nuklearnom megnetskom rezonancijom u čvrstom stanju (engl. *solid-state NMR*),
- elektronskom kristalografijom (engl. *electron crystallography*),
- elektronskom mikroskopijom (engl. *electron microscopy*), te
- difrakcijom X-zraka na vlaknastim proteinskim strukturama (engl. *X-ray fiber diffraction*).

Početni razmatrani skup TM proteina alfa vrste S481 sadrži 481 proteinski lanac [41], a u tom je radu razmatran i njegov podskup S392 s 392 lanca (u radu nazvan i standardni skup). Taj skup uključuje samo lance čije su strukture riješene metodom rendgenske difrakcije (uz rezoluciju do 3.5Å), ili nuklearnom magnetskom rezonancijom u otopini. Za analize tih skupova proteinskih lanaca napravljen je i poslužitelj dostupan putem interneta [41]. Na njemu se mogu uspoređivati metode za predviđanje topologije i topografije proteinskih lanaca, ali i izabrati reprezentativne skupove proteinskih lanaca za razine pragove sličnosti ili identičnosti. U analizama sličnosti mogu se izabrati dva algoritma za sravnjenja proteinskih slijedova (Needleman-Wunsch za globalno sravnjenje, i Smith-Waterman za lokalno sravnjenje) uporabom programa EMBOSS [49].

Skupovi S481 i S392 korišteni su u razvoju i analizi algoritama za izbor reprezentativnih skupova membranskih proteina razvijenih u disertaciji, i za njihovu usporedbu s algoritmom Hobohm 2 [24] koji su izradili autori rada [41] i postavili ga na mrežni poslužitelj. Zapisi proteinskih slijedova u skupovima S481 i S392 sadrže primarnu strukturu (aminokiselinski slijed), informacije o topologiji, te sekundarnu strukturu, tj. položaje TM segmenata u strukturi. Nakon toga se skriptom napisanom u programskom jeziku Python naziva S481_topol-koef-OPM.py (Prilog 1e disertaciji, 'e' je oznaka za elektronički) iz dobivenih slijedova i informacija o strukturi, te iz baze podataka OPM pronalaze, filtriraju i preuzimaju podaci o:

(1) eksperimentalnoj metodi kojom je riješena struktura,
(2) rezoluciji i kratkom opisu proteina,
(3) broju TM segmenata (i podaci o njihovim položajima i duljinama),
(4) ukupnom broju aminokiselinskih ostataka unutar i izvan membrane.

Najvažniji podaci za skup S481 ukratko su prikazani u tablici 3.

Tablica 3. Osnovne osobine skupa S481 iskazane po lancu i za cijeli skup.

skup S481	AK	TM segment	AK u membrani	% AK u membrani
minimum po lancu	25	1	17	2.40
maksimum po lancu	1278	20	417	79.31
srednja vrijednost po lancu	282.27	4.86	113.66	40.49
ukupno – skup	135770	2337	54670	

AK – aminokiselina, TM – transmembranski

Vidi se da su duljine slijedova u rasponu od 25 do 1278 aminokiselina, a raspon broja TM segmenata u proteinskim lancima je od 1 do 20. Prosječni je broj TM segmenata po lancu 4.86, a njihov ukupni broj u skupu iznosi 2337.

Iz skupa S481 autori su dobili pročišćeni skup S392 s 392 proteinska lanca koji je u radu [41] bio polazišni skup za primjenu njihove inačice algoritma Hobohm 2 za izbor reprezentativnog skupa. U konačnici, izdvojen je reprezentativni skup sa 101 lancem međusobne

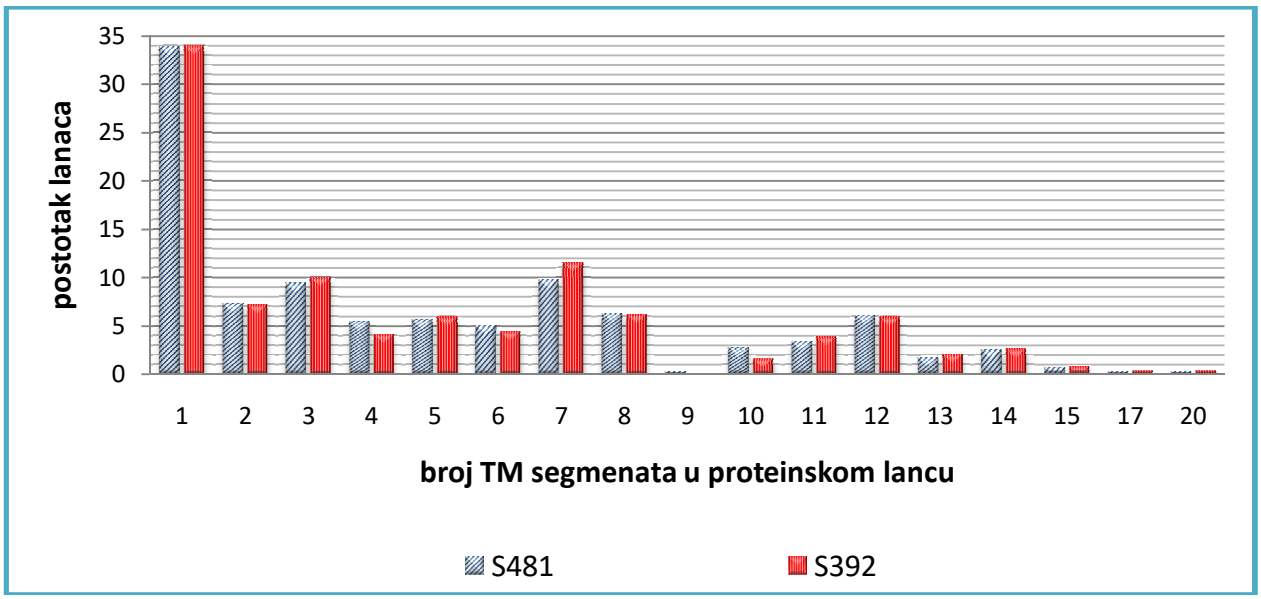
sličnosti ispod 30% [41], što je bio jedan od važnijih doprinosa u tom radu. Analizom aminokiselinskih slijedova i struktura dobiveni su osnovni podaci za taj skup (tablica 4). Usporedba s početnim skupom S481 (tablica 3) pokazuje da podskup S392 ima prosječno malo duže proteinske lance s nešto više TM segmenata po lancu. To ukazuje na to da reprezentativni podskup S392 sadrži (prosječno) nešto složenije strukture.

Tablica 4. Osnovne osobine skupa S392 iskazane po lancu i za cijeli skup.

skup S392	AK	TM segment	AK u membrani	% AK u membrani
minimum po lancu	25	1	18	2.40
maksimum po lancu	1051	20	417	79.31
srednja vrijednost po lancu	279.24	4.90	115.26	40.70
ukupno – skup	109462	1921	45183	

AK – aminokiselina, TM – transmembranski

Kako bi se uvidjela glavna obilježja ovih početnih skupova s obzirom na broj TM segmenata u lancima, napravljena je raspodjela broja lanaca (u postotcima) u ovisnosti o broju TM segmenata u lancima (slika 7).



Slika 7. Postotak proteinskih lanaca u početnim skupovima S481 i S392 po podskupovima lanaca istog broja TM segmenata.

Vidi se da je trećina lanaca u oba skupa S481 i S392 s jednim TM segmentom, a znatnije su zastupljeni lanci s 3 i 7, te s 2, 8 i 12 TM segmenata. Sve te skupine lanaca imaju više od 5% ukupnog broja lanaca u skupovima. Zanimljivo je da skoro potpuno izostaju lanci s devet TM segmenata, tj. među riješenim strukturama u bazi PDB mali je broj takvih lanaca. Međutim, možda je među integralnim membranskim proteinima u živom svijetu općenito mali broj struktura s devet TM segmenata.

Primjena algoritma Hobohm 2 na početni skup S481 iz [41]

Radi što kvalitetnije usporedbe algoritama, provedeni su izbori reprezentativnih skupova proteina polazeći svaki put od skupa S481 za različite razine (pragove) sličnosti i identičnosti u granicama od 20% – 35%.

Bliskost ili udaljenost proteinskih slijedova (lanaca) može se iskazati usporedbom odgovarajućih mjesta u sravnjenju aminokiselinskih slijedova promatrajući pritom:

(a) identičnost aminokiselina (potpuna podudarnost) u oba slijeda na istim mjestima u sravnjenju (što nazivamo '*identičnošću*' i iskazujemo u postotcima), ili

(b) potpunu podudarnost aminokiselina ali i sličnost (srodnost) aminokiselina tamo gdje nema podudarnosti (identičnosti) (što nazivamo '*sličnošću*' i iskazujemo u postotcima).

Kod sravnjenja u kojima se kvantificira bliskost/udaljenost među aminokiselinskim slijedovima preko identičnosti, broje se samo identične aminokiseline na odgovarajućim položajima (mjestima) u prvom i drugom lancu u sravnjenju. U sravnjenjima koja kvantificiraju sličnost osim doprinosa identičnih aminokiselina na odgovarajućim položajima, zbrajaju se i doprinosi ne-identičnih aminokiselina. Doprinosi ne-identičnih aminokiselina u sravnjenju imaju različite iznose, a ti iznosi iskazni su odabranom matricom sličnosti aminokiselina. Postoje brojne matrice sličnosti, a temelje se npr. na nekom svojstvu aminokiselina, ili na matrici učestalosti (dopustivih) zamjena jedne aminokiseline nekom drugom u proteinima promatrano tijekom evolucije. Matrica sličnosti odabire se u samom početku rada programa EMBOSS [49] rabljenog u disertaciji.

Primjenom inačice algoritma Hobohm 2 razvijene u'grupi Sydney' [41] na skup S481, dobivaju se vrijednosti prikazane u tablici 5 [41].

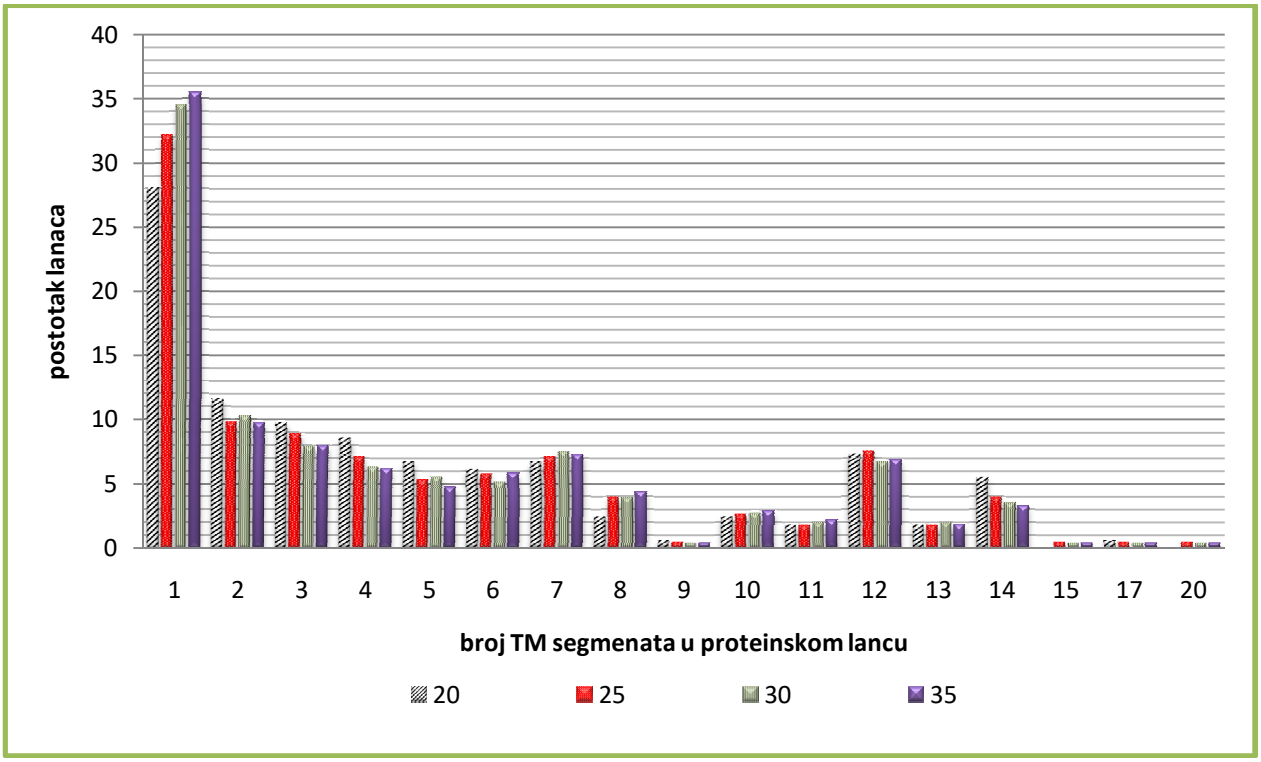
Tablica 5. Osobine reprezentativnih skupova izabranih algoritmom Hobohm 2 iz skupa S481 za pragove identičnosti i sličnosti u rasponu 20% – 35%.

identičnost [a]			
razina/prag (%)	broj lanaca	broj TM segmenata	broj AK
20	164	820	47335
25	224	1109	62670
30	252	1205	67187
35	276	1309	74170
sličnost [a]			
razina/prag (%)	broj lanaca	broj TM segmenata	broj AK
20	32	129	9644
25	63	255	19882
30	121	580	37162
35	178	903	50237

[a] Podaci se odnose za cijeli skup; Kratice: TM –transmembranski, AK – aminokiselina

Ako usporedimo broj lanaca ili TM segmenata u reprezentativnim skupovima za odgovarajuće (iste) razine sličnosti i identičnosti (tablica 5), vidimo da je rezultat po identičnosti na razini 20% najbliži rezultatu za sličnost na razini 35%. Također, skup izabran uz prag identičnosti 20% osjetno je veći, ali podizanjem praga porast broja lanaca (ili TM segmenata) sporiji je nego u slučaju analize sličnosti, gdje je ukupni broj lanaca kod praga 35% 5.6 puta veći nego kod praga sličnosti od 20%. Očigledno je da je strogost (restriktirnost) izbora kad se koristi identičnost osjetno manja nego kada se koristi sličnost među slijedovima.

Raspodjela broja lanaca (iskazana u postotnim udjelima) s određenim brojem TM segmenata u strukturi za četiri reprezentativna skupa, prikazane su na slici 8. Ti su skupovi izabrani polazeći od početnog skupa S481 za četiri praga identičnosti (20, 25, 30 i 35%).

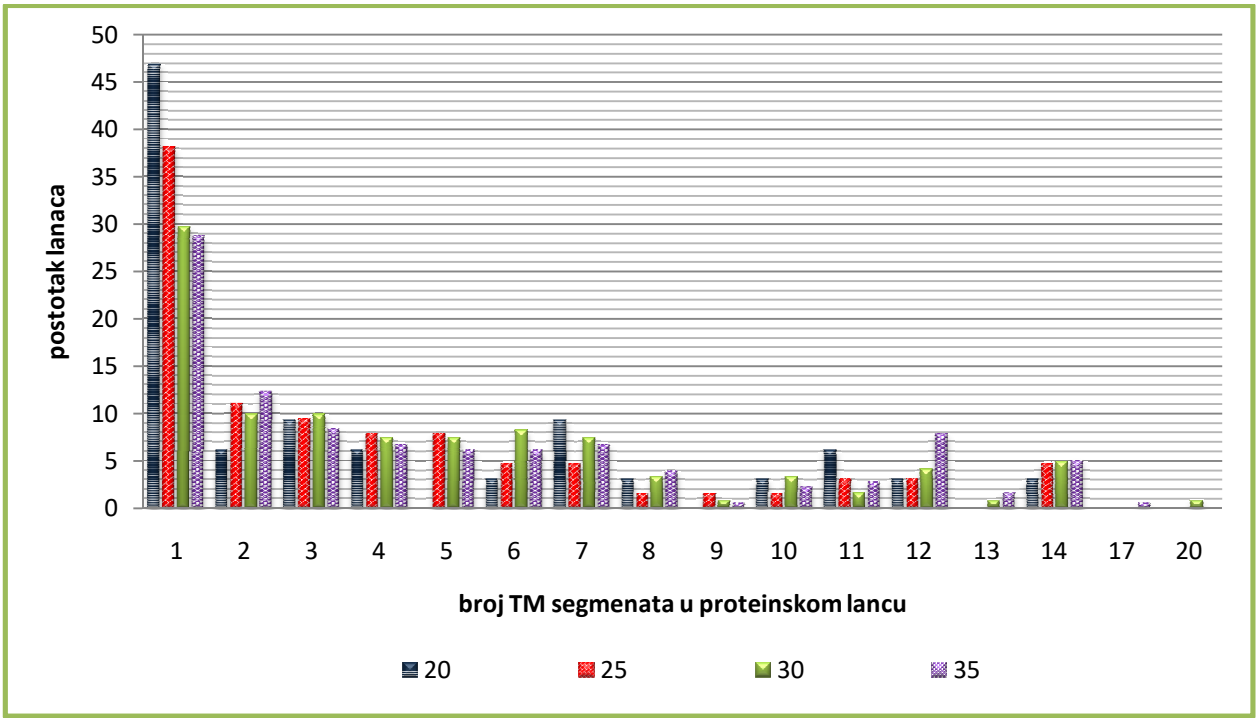


Slika 8. Raspodjele broja lanaca (u postotcima) po podskupovima lanaca istog broja TM segmenata u reprezentativnim skupovima dobivenim od početnog skupa S481 algoritmom Hobohm 2 za različite pragove identičnosti.

Na granici 20% identičnosti postotni udio lanaca s jednim TM segmentom doseže 28%. Kako se prag identičnosti povećava, udio tih proteina raste do približno 36%. Istovremeno, broj proteinskih lanaca koji imaju od 2 do 5 TM segmenata opada s povećanjem praga identičnosti do 35%, nadomještajući tako značajniji porast broja lanaca u skupini proteina s jednim TM segmentom (slika 8).

Porastom praga sličnosti između lanaca (čime kriterij izbora postaje manje strog), postotni udio lanaca s jednim TM segmentom opada do 28% kod praga sličnosti od 35% (slika 9). Ovaj trend u analizi identičnosti kod proteina koji imaju jedan TM segment zanimljiv je i suprotan trendu kad se u analizama računa sličnost. U izboru reprezentativnog skupa prema sličnosti vidi se da, ako je prag 20%, postotni udio lanaca s jednim TM segmentom doseže skoro polovicu (47%) ukupnoga broja lanaca.

Podizanjem praga dopustive sličnosti među lancima (čime kriterij izbora postaje manje strog), postotni udio lanaca s jednim TM segmentom opada do 28% kod praga sličnosti od 35% (slika 9). Pad broja proteinskih lanaca s jednim TM segmentom prati osjetniji porast broja lanaca s 2, 6, 12 i 13 TM segmenata. Pritom treba imati na umu kako su proteinski lanci s više TM segmenata (u pravilu) i dulji.



Slika 9. Raspodjele broja lanaca (u postotcima) po podskupovima lanaca istog broja TM segmenata u reprezentativnim skupovima dobivenim od početnog skupa S481 algoritmom Hobohm 2 za različite pragove sličnosti.

Primjena algoritma Hobohm 2 na početni skup S392 iz [41]

U nastavku bit će prikazani analogni rezultati izbora reprezentativnih skupova proteina na razinama sličnosti i identičnosti u granicama od 20% – 35% uporabom algoritma Hobohm 2 primjenjenog od 'grupe Sydney' [41]. Pritom, svaki put polazi se od standardnog skupa S392, za koji su u radu [41] prikazani najvažniji rezultati (tablica 6).

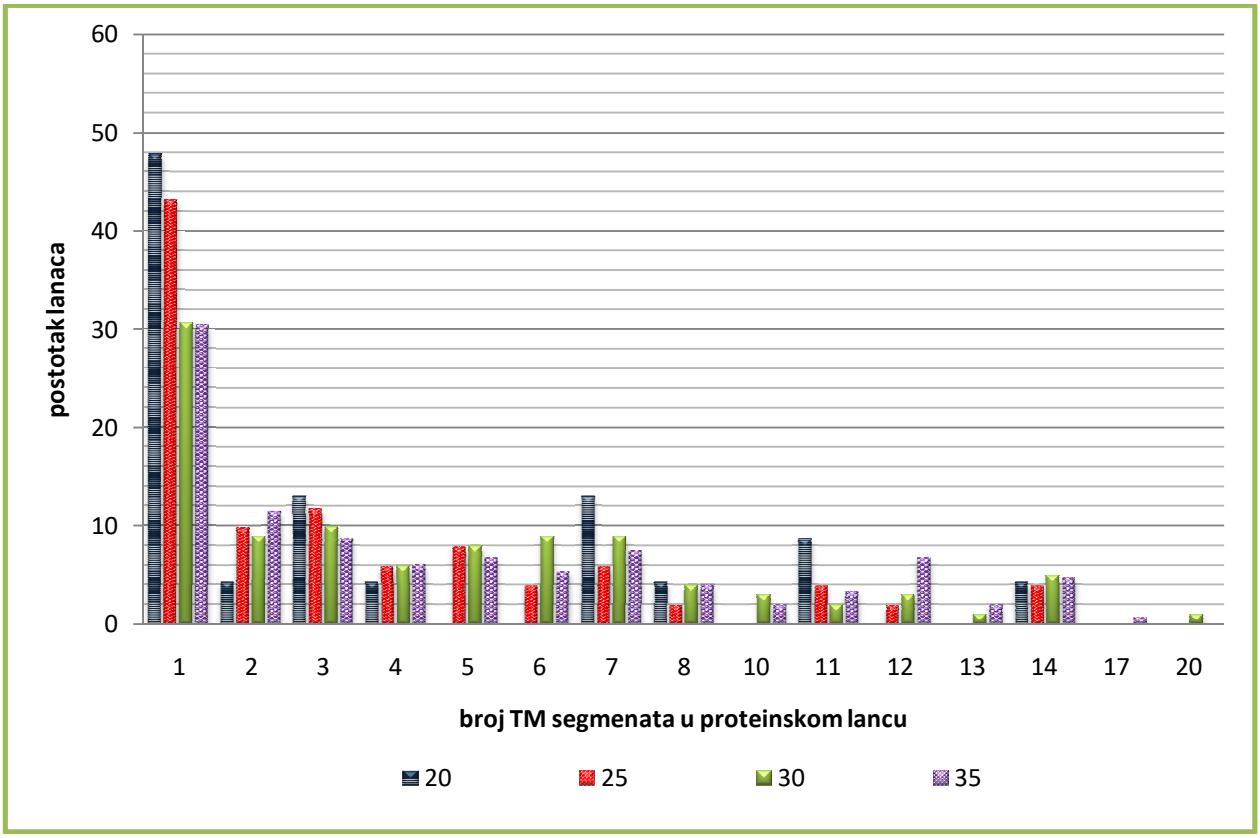
Tablica 6. Osobine reprezentativnih skupova izabranih algoritmom Hobohm 2 iz skupa S392 za pragove identičnosti i sličnosti u rasponu 20% – 35%.

identičnost			
razina/prag (%)	broj lanaca	broj TM segmenata	broj AK
20	134	638	36200
25	183	917	50447
30	208	995	53960
35	227	1075	58974
sličnost			
razina/prag (%)	broj lanaca	broj TM segmenata	broj AK
20	23	91	6195
25	51	185	13998
30	**101**	**483**	**30144**
35	148	736	39642

[a] Podaci se odnose za cijeli skup iz ref. [41][a]; Kratice: TM –transmembranski, AK – aminokiselina

Podebljanim brojevima u tablici 6 označen je standardni skup na razini 30% sličnosti, koji je u radu 'grupe Sydney' [41] bio prikazan kao glavni reprezentativni skup dobiven njihovom

izvedbom algoritma Hobohm 2 [25] iz početnog skupa S392. Ovi rezultati važni su i stoga što će glavni rezultati s algoritmima razvijenima u disertaciji biti dobiveni polazeći od skupa S392. Kvaliteta tih glavnih rezultata (reprezentativnih skupova) ocjenjivat će se i uspoređivati brojem lanaca i/ili brojem TM segmenata, te kompleksnošću proteinskih struktura lanaca izabranih u reprezentativni skup.



Slika 10. Raspodjele broja lanaca (u postotcima) po podskupovima lanaca istog broja TM segmenata u reprezentativnim skupovima dobivenim od početnog skupa S392 algoritmom Hobohm 2 za različite pragove sličnosti.

U izborima reprezentativnih skupova polazeći od podskupa S392 (kao i u slučaju kad se polazilo od skupa S481), porastom granice sličnosti između lanaca opada postotni udio lanaca s jednim TM segmentom. Za prag sličnosti 20% udio lanaca s jednim TM segmentom skoro doseže 50% ukupnog broja lanaca (slika 10), dok za prag 35% taj udio padne na 30%.

Polazeći od početnog skupa S392 uz prag sličnosti 30% uz korištenje njihove izvedbe algoritma Hobohm 2 [41], autori su dobili standardni reprezentativni skup niske zalihosti (niske međusobne sličnosti) sa 101 proteinskim lancem označen u disertaciji kao S101. Osnovna obilježja skupa S101 prikazana su u tablici 7.

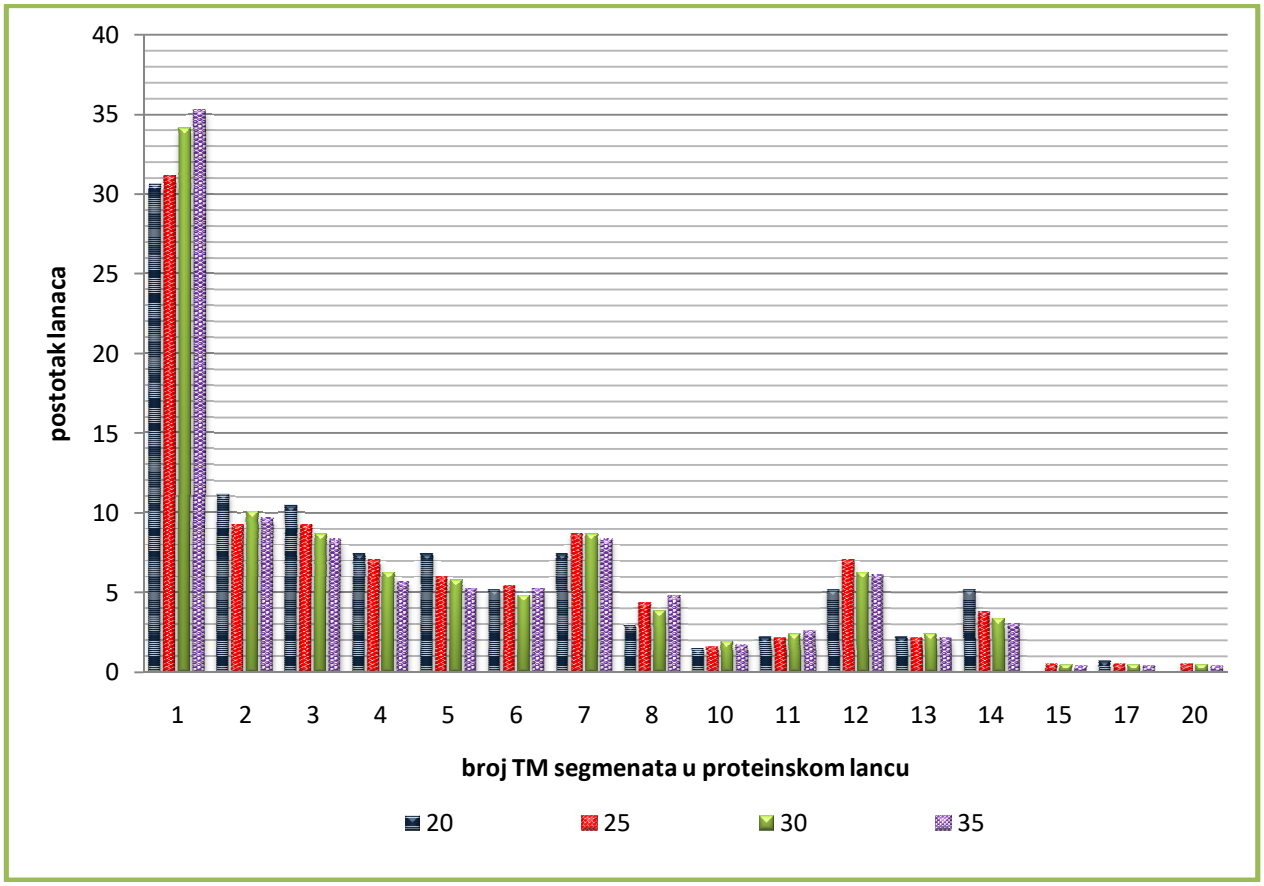
Tablica 7. Osnovne osobine reprezentativnog skupa S101 iskazane po lancu i za cijeli skup.

skup S101	AK	TM segment	AK u membrani	% AK u membrani
minimum po lancu	31	1	20	6.98
maksimum po lancu	1047	20	417	77.42
srednja vrijednost po lancu	298.46	4.78	113.45	35.74
ukupno – skup	30144	483	11458	

AK – aminokiselina, TM – transmembranski

Na slici 10. (označeno zelenom bojm, prag 30%) vidi se da je i u reprezentativnom skupu S101 udio lanaca s jednim TM segmentom poprilično visok, i to oko 31%. Međutim, kako početni skup sadrži skoro polovicu lanaca s jedim TM segmentom, vidi se da je algoritam optimiran kako bi birao u reprezentativni skup više lanaca koji sadrže više od jednog TM segmenata (a manje lanaca s jednim TM segmentom).

Nadalje, rezultati za reprezentativne skupove koji se izabiru istim algoritmom (izvedbom algoritma Hobohm 2 u ref. [41]) ali za različite razine identičnosti između proteinskih slijedova, prikazani su na slici 11.



Slika 11. Raspodjele broja lanaca (u postotcima) po podskupovima lanaca istog broja TM segmenata u reprezentativnim skupovima dobivenim od početnog skupa S392 algoritmom Hobohm 2 za različite pragove identičnosti.

Uočava se da broj lanaca u tim skupovima s jednim TM segmentom raste s porastom razine identičnosti (veći broj lanaca za prag identičnosti 35%). To je u suprotnosti s rezultatima analize za različite razine (pragove) sličnosti jer je u tom slučaju broj lanaca s jednim TM segmentom opadao (slika 10). Usto, porastom praga identičnosti od 20% do 35% uočava se pad postotnog udjela lanaca s 3, 4, 5 i 14 TM segmenata.

2.4.3. Opis skupa M1087 (i njegovog podskupa M166)

U radu "grupe München" [43] izdvojen je skup TM proteina poznate strukture s TM segmentima u sekundarnoj strukturi alfa uzvojnice, polazeći od zapisa proteina u bazama podataka OPM i PDBTM iz srpnja 2013. godine. Obje baze koriste PDB identifikatore (kodove) lanaca, koji su dodatno povezani s njihovim UniProtKB kodovima. Na taj način izdvojeno je 1087 proteinskih

lanaca iz (ukupno) 455 struktura TM proteina pohranjenih u PDB (379 određenih rentgenskom difrakcijom rezolucije ≤ 8Å i 76 struktura određenih s pomoću NMR metode u otopini.

Početnom analizom proteinskih lanaca skupa M1087 ('M' od 'München', a 1087 broj je lanaca u skupu) uočen je veliki broj međusobno potpuno identičnih slijedova. Nakon izuzimanja identičnih lanaca (programski su izdvojeni samo slijedovi koji se međusobno razlikuju barem u jednoj aminokiselini), od početnih 1087 lanaca preostaje njih samo 430 za daljnju analizu i izbor reprezentativnog skupa. Osnovna obilježja skupa M1087 dobivenog od autora rada [43] daje podatke prikazane u tablici 8.

Tablica 8. Osnovne osobine skupa M1087 iskazane po lancu i za cijeli skup.

skup M1087	AK	TM segment	AK u membrani	% AK u membrani
minimum po lancu	29	1	6	1.67
maksimum po lancu	1342	19	443	79.31
srednja vrijednost po lancu	320.63	4.70	100.46	37.94
ukupno – skup	348530	5106	109197	

AK – aminokiselina, TM – transmembranski

Koristeći algoritam UniqueProt [44] „grupa München“ izabrala je reprezentativni skup sa 166 proteinskih lanaca s obilježjima danim u tablici 9. Usporedbom početnog skupa M1087 i reprezentativnog skupa M166 izabranog uporabom algoritma UniqueProt uz prag identičnosti 20%, uočava se manje TM segmenata po proteinskim lancu (3.25 u odnosu na 4.7 u M1087), kao i (posljedično) manji prosječni broj aminokiselina u TM segmentima po lancu.

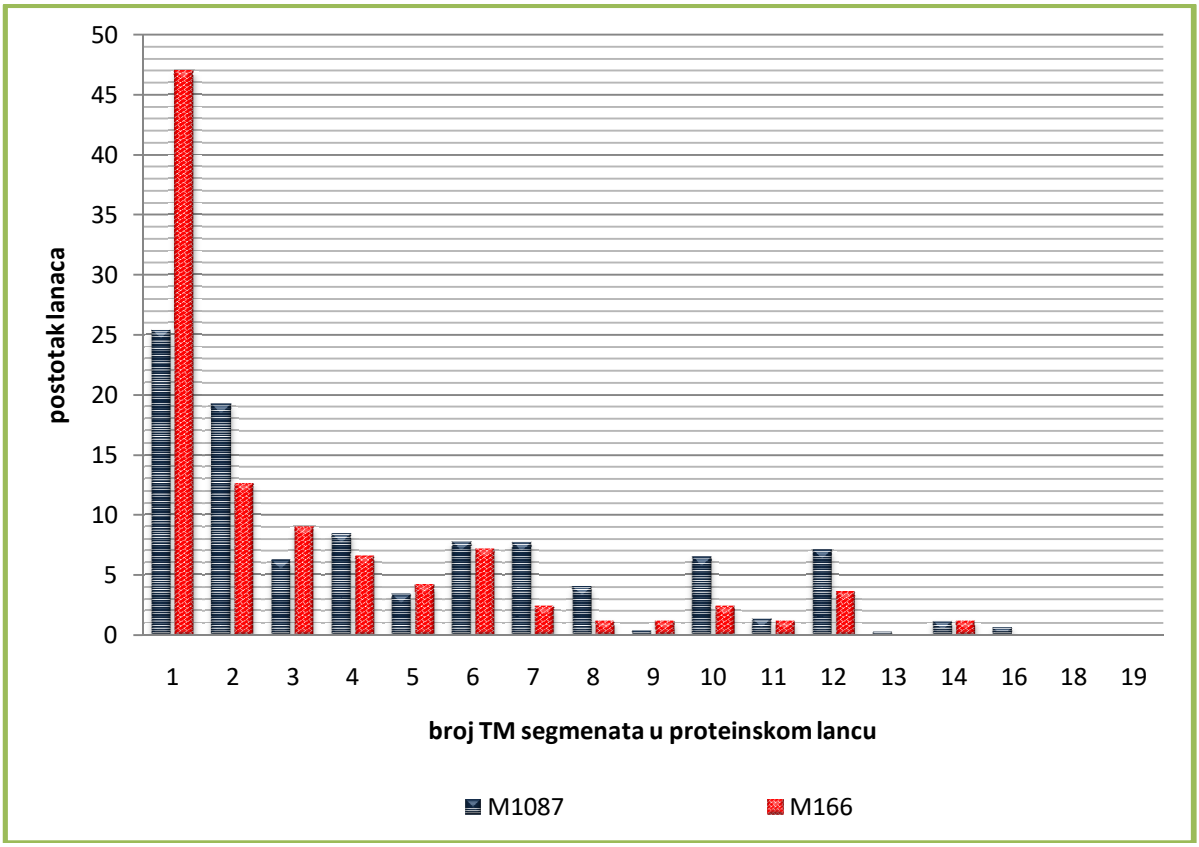
Tablica 9. Osnovne osobine reprezentativnog skupa M166 iskazane po lancu i za cijeli skup.

skup M166	AK	TM segment	AK u membrani	% AK u membrani
minimum po lancu	29	1	6	2.12
maksimum po lancu	1001	14	299	79.31
srednja vrijednost po lancu	200.61	3.25	69.48	39.61
ukupno – skup	33301	540	11533	

AK – aminokiselina, TM – transmembranski

Prosječni broj TM semenata u reprezentativnom skupu M166 manji je za više od 30% u odnosu na početni skup M1087. Ta je razlika na opisanim odgovarajućim skupovima „grupe Sydney“ S392 i S101 znatno manja, i iznosi približno 2.5%. Iz toga zaključujemo da algoritam UniqueProt [44] provodi izbor reprezentativnoga skupa i smanjenje identičnosti tako da u zadržavanju u skupu daje prednost kraćim lancima, s manje TM segmenata.

Usporedba udjela lanaca s istim brojem TM segmenata u početnom skupu M1087 i u reprezentativnom skupu M166 izabranom metodom UniqueProt [44] (iz M1087) za prag identičnosti 20%, ukazuje na značajno povećanje postotnog udjela lanaca s jednim TM segmentom (slika 12). Naime, udio lanaca s jednim TM segmentom promijeni se s 25% u početnom skupu M1087, na 47% u izabranom reprezentativnom skupu M166. Detaljniji prikaz dan je na slici 12, gdje se vidi i to da je broj lanaca s 2,7,8,10 i 12 TM segmenata u osjetnome padu u skupu M166.



Slika 12. Postotak proteinskih lanaca u početnom skupu M1087 i reprezentativnom skupu M166 u ovisnosti o broju TM segmenata.

Ovakav rezultat navodi nas na zaključak da, možda u nastojanju za izborom što većeg broja lanaca u reprezentativnom skupu, algoritam UniqueProt učestalije bira lance s manjim brojem TM segmenata, koji su bitno jednostavnije strukture i topologije. Uzimajući u obzir evolucijsku informaciju koja ukazuje na to da su funkcionalno i strukturno bitniji dijelovi lanca (npr. TM segmenti) bolje očuvani nego preostali dijelovi, može se izvući zaključak kako su veći dijelovi aminokiselinskog slijeda bolje očuvani u proteinskim lancima s više TM segmenata. Posljedično, lanci s više TM segmenata zadržavaju i značajniju međusobnu identičnost (i/ili sličnost). Suprotno tome, kod lanaca s jednim TM segmentom evolucijski bolje očuvani dio lanca samo je u predjelu oko tog jednog TM segmenta koji predstavlja manji dio lanca. Stoga, takvi proteinski lanci pokazuju nižu međusobnu identičnost (i/ili sličnost). Ovome u prilog idu sljedeći podaci dobiveni analizom skupa M1087:

(1) s jednim TM segmentom ima 276 lanaca sa 63546 aminokiselinskih ostataka (prosječno 230.24 po lancu),
(2) s dva TM segmenta ima 209 lanaca s 40025 aminokiselina (prosječno 191.51 po lancu),
(3) s tri TM segmenta ima 68 lanaca s 14241 aminokiselina (prosječno 209.43 po lancu),
(4) tek lanci s četiri TM segmentai s prosječno 323.30 aminokiseline po lancu prelaze prosječnu duljinu onih lanaca s jednim TM segmentom.

2.4.4. Opis skupa S148

Skup S148 podskup je skupa S392 (skup "grupe Sydney" [41]) dobiven odabirom lanaca jednom ranom inačicom algoritma razvijenog u disertaciji. Tim se algoritmom nastojalo u reprezentativnome skupu zadržati one proteine koji imaju:

(1) veći broj TM segmenata, i (ako taj nije bio dovoljan za odluku)

(2) strukturu riješenu s nižom (boljom) rezolucijom (tamo gdje nisu korištene mogućnosti nasumičnog odabira lanca, niti dodatna sortiranja pri iteraciji).

Skup S148 kreiran je radi usporedbe metoda razvijenih u disertaciji s metodom "grupe München". Ta je grupa koristila pri izboru reprezentativnih skupova u radovima [42,43] dvije inačice algoritma UniqueProt [44]. Na taj se način nastojala provjeriti kvaliteta i strogost algoritama u međusobnim unakrsnim usporedbama. Osnovna obilježja izabranoga skupa S148 prikazane su u tablici 10.

Tablica 10. Osnovne osobine skupa S148 iskazane po lancu i za cijeli skup.

skup S148	AK	TM segment	AK u membrani	% AK u membrani
minimum po lancu	25	1	16	3.85
maksimum po lancu	1051	20	370	78.13
srednja vrijednost po lancu	283.55	5.10	105.41	39.19
ukupno – skup	41966	755	15601	

AK – aminokiselina, TM – transmembranski

Važno je uočiti veliki prosječni broj od 5.1 TM segmenata po lancu, što je osjetno više nego u reprezentativnome skupu M166 koji je slične veličine a ima prosječno 3.25 TM segmenata po lancu (tablica 9). Detaljnija usporedba i analiza različitih algoritama na skupu S148 izložena je u dijelu rezultati.

2.4.5. Opis skupova N189 i N263

Analizama cijele baze podataka membranskih proteina OPM koji su u nju bili pohranjeni do početka 2017. godine, izdvojen je početni skup N1043 ('N' je oznaka za 'našu' inačicu skupa) s 1043 proteinska lanca s TM segmentima alfa sekundarne strukture. Primjenom algoritma 1 (opisanog u poglavlju 2.5) razvijenog u disertaciji, nastojalo se u reprezentativnim skupovima zadržati proteine koji imaju: (1) veći broj TM segmenata i, ako taj nije bio dovoljan za odluku, (2) strukturu riješenu s nižom (boljom) rezolucijom. Tako su dobiveni reprezentativni skupovi N189 i N263 za potrebe usporedbi algoritama iz disertacije s algoritama iz literature. Skup N189 izabran je uz prag identičnosti 18%, a skup N263 uz prag identičnosti 20%. Matrice sličnosti/identičnosti između početnih proteinskih slijedova na temelju kojih algoritmi izabiru reprezentativne skupove, dobivene su programom EMBOSS [49].

Tablica 11. Osnovne osobine skupa N189 iskazane po lancu i za cijeli skup.

skup N189	AK	TM segment	AK u membrani	% AK u membrani
minimum po lancu	29	1	10	2.39
maksimum po lancu	1321	17	341	81.75
srednja vrijednost po lancu	324.61	4.98	105.09	36.24
ukupno – skup	61352	942	19862	

AK – aminokiselina, TM – transmembranski

Na taj se način nastojalo provjeriti koliko proteinskih lanaca otpada iz reprezentativnog skupa ako se prag identičnosti spusti malo (za 2%) ispod 20%, tj. najčešće korištene vrijednosti za prag identičnosti.

Skup N189 izabran je i uz dodatni kriterij prema kojemu se nastoji u svakom koraku u skupu ostaviti lanac s najmanjim brojem susjeda (uz nasumičnu preraspodjelu lanaca s

najmanjim brojem susjeda u svakoj iteraciji). Skup N263 dobiven je bez dodatnih uvjeta, tj. nisu korištene mogućnosti nasumičnog odabira lanca, a niti dodatno sortiranje/redanje pri kojemu bi se ostavljao lanac s najmanjim brojem susjeda (primjenom algoritma 1). Osobine skupa N189 (dobiven uz prag identičnosti 18%) prikazane su u tablici 11, a skupa N263 (dobivenoga uz prag identičnosti 20%) u tablici 12.

Tablica 12. Osnovne osobine skupa N263 iskazane po lancu i za cijeli skup.

skup N263	AK	TM segment	AK u membrani	% AK u membrani
minimum po lancu	25	1	11	2.12
maksimum po lancu	1321	19	443	84.00
srednja vrijednost po lancu	310.16	5.56	117.29	40.43
ukupno – skup	81572	1461	30847	

AK – aminokiselina, TM – transmembranski

Uočavaju se veliki prosječni brojevi TM segmenata po lancu (4.98 i 5.56) za skupove N189 i N263 u odnosu na druge skupove analizirane u radu. Također, vidimo da se u spuštanju praga identičnosti za samo 2% broj lanaca u izabranom skupu smanji za skoro 30%, a broj TM segmenata za 35%. Detaljnija analiza rezultata dobivenih na skupovima N189 i N263 različitim algoritmima, kao i dobiveni reprezentativni skupovi, biti će prikazani u poglavlju rezultati.

3. REZULTATI I RASPRAVA

S obzirom da je temeljna istraživačka zadaća disertacije metodološka - razvoj algoritma i unapređenje metodologije za izbor reprezentativnog skupa membranskih proteina, te u sklopu toga i pronalaženje teorijske osnove i interpretacije novo-uvedenih kriterija koji kvantificiraju složenost modelne strukture proteina, taj dio rezultata dobivenih u disertaciji bit će opisan u prvom dijelu rezultata (pod-poglavlje 3.1).
Novo-uvedeni koncepti i razvijeni algoritmi prikazani su i raspravljeni u odjeljcima:

a) 3.1.1, gdje je uveden koncept nasumičnog modela za strukturu s dva stanja i izveden izraz za izračun najvjerojatnije točnosti nasumičnog modela ($Q_{2,rnd}$) na primjeru membranskih proteina,
b) 3.1.3, gdje je definiran i uveden koncept binomnog nasumičnog modela za strukturu s dva stanja, i izveden izraz za izračun broja realizacija modelne strukture membranskih proteina,
c) 3.1.4 gdje je po prvi put uveden koncept segmentnog nasumičnog modela i izvedeni izrazi za izračun broja realizacija modelne strukture za segmente različitih duljina, te za slučaj kad između segmenata ne postoji (ili kad postoji) minimalni razmak,
d) 3.1.5 gdje je dana fizikalna interpretacija rezultata vezom između broja realizacija modelne strukture i entropije,
e) 3.1.6 i 3.1.7 gdje su razvijeni Algoritmi 1, 2 i 3 za provedbu izbora repezentativnih skupova integralnih membranskih proteina alfa vrste.

Drugi dio rezultata prikazan je u pod-poglavljima:

a) 3.2. gdje su u ovoj disertaciji razvijeni Algoritmi 1, 2 i 3 uspoređeni s rezultatima i metodama drugih istraživača [24,25,39,40,41,46] na nekoliko skupova proteinskih slijedova (lanaca) poznate strukture,
b) 3.3. gdje su prikazani najnoviji (bitno veći) reprezentativni skupovi integralnih membranskih proteina alfa vrste (niske međusobne sličnosti) izabrani Algoritmima 1, 2 i 3, polazeći od najnovijih inačica baza PDB [20] i OPM [29] iz 2017. godine,
c) 3.4. gdje su uspoređeni rezultati dobiveni različitim metodama i na različitim skupovima.

3.1. Teorijske metode i računalni algoritmi razvijeni i korišteni u disertaciji

U analizama postojećih algoritama za izbor reprezentativnog skupa (Hobohm 1, Hobohm 2 i UniqueProt) uočene su mogućnosti poboljšanja izbora uvođenjem:

(1) nasumičnog odabira lanaca (neovisnost o ulaznom poretku), ali isto tako i

(2) odabira lanaca složenije strukture u konačni reprezentativni skup. Jedan od razloga je u tome što bi skup proteinskih lanaca složenije strukture bio zahtjevniji za metode koje se koriste za predviđanja položaja TM segmenata.

S ciljem iskazivanja (kvantificiranja) složenosti neke strukture (tj. membranskog proteina) razvijen je koncept nasumičnog modela uz korištenje parametra točnosti Q_2 koji se uobičajeno navodi u objavljenim radovima. Nadalje, kako bi se procijenila maksimalna konformacijska entropija modelne strukture membranskog proteina, uveden je i koncept nasumičnog modela kojim se računa maksimalni broj mogućih realizacija modelne strukture (odnosno broj modelnih konformacija).

3.1.1. Koncept nasumičnog modela ($Q_{2,rnd}$ parametar)

Prilikom analiza i procjene kvalitete modela za predviđanje sekundarne strukture proteina kad su prisutne samo dvije vrste sekundarne strukture, najjednostavniji parametar je Q_2. Ovaj parametar iskazuje točnost modela u predviđanju oba stanja sekundarne strukture membranskih proteina.

Parametar Q_2 naziva se u literaturi još i klasifikacijska točnost modela u problemu koji ima definirana dva stanja [65], ili kao postotak svih ispravnih predviđanja [66]. Ako definiramo jedno stanje kao dio strukture i proteinskog slijeda koji je 'u membrani', onda je drugo stanje dio strukture i proteinskog slijeda koji je 'izvan membrane'. Nastoji se u disertaciji razviti klasifikacijski model s dva stanja koji za odabrano svojstvo ili aktivnost (Y-varijablu) pomoću jednog ili više ulaznih molekularnih atributa (deskriptora, tj. X-varijabli) koji, do određene točnosti, ispravno procjenjuje ili predviđa strukturnu klasu ili svojstvo.
Neka je prema modelu (razvijenom na skupu s dvije klase koje imaju ukupno *N* elemenata) broj ispravno klasificiranih elemenata prve klase jednak *p* a druge klase *n*, parametar Q_2 je:

$$Q_2 = \left(\frac{p+n}{N} \cdot 100\right)(\%)$$

Ako se za klasifikacijski model s dva stanja dobije vrijednost parametra Q_2 od 90% (ili 95%), čini se da je model iznimno (i impresivno) točan. Međutim, stvarna razina točnosti modela (odnosno njegovog doprinosa) može se procijeniti tek ako se ta Q_2 vrijednost uspoređuje s točnošću koja se može dobiti nasumičnim modelom $Q_{2,rnd}$ (odnosno, s najvjerojatnijim nasumičnim modelom). Očigledno je stvarni doprinos modela u gore spomenutom slučaju ($Q_2 = 90\%$) značajno različit ako je najvjerojatnija nasumična točnost $Q_{2,rnd} = 50\%$, ili ako je $Q_{2,rnd} = 70\%$. Za svaki model, a također i za modele odnosa strukture i svojstava molekula koji se odnose na male molekule ili proteine, moguće je izračunati (ili procijeniti simulacijama) razinu točnosti koja se može dobiti nasumičnim modelom. Pritom, u procjeni točnosti nasumičnoga modela moguće je koristitinasumično presložene (sortirane) izvorne podatke (varijable), ili koristiti nasumične podatke (varijable) dobivene generatorima nasumičnih brojeva. Kada je točnost modela iskazana (procijenjena) nekim statističkim parametrom, druga važna vrijednost koju je potrebno izračunati i dati uz model je vrijednost istog statističkog parametra za odgovarajući nasumični model.

Iako su dva rada koja su se bavila ovom temom u analizi koeficijenta korelacija modela objavljeni još prije četrdesetak godina za multivarijantne linearne regresijske modele [67,68], preporuke iz tih radova u redovitoj su primjeni u postupcima modeliranja. U tim je radovima provedena analiza nekoliko modela kako bi se došlo do informacije o maksimalnom prihvatljivom broju varijabli u modelima multivarijante linearne regresije (MLR) u ovisnosti o veličini skupa podataka. Kao rezultat analiza proizašla je procjena (preporuka) da broj varijabli (deskriptora) uključenih u MLR modele ne bi smio prelaziti 1/5 broja slučajeva (molekula) u skupu podataka [67].

Nasumična korelacija (ili točnost) veća je za stvarne nego za nasumične parove varijabli, jer stvarne varijable imaju (obično) monotonije raspodjele vrijednosti od nasumičnih. Osim toga, stvarne varijable imaju, u pravilu, temeljnu korelaciju s nekim osnovnim svojstvima skupa podataka. U slučaju skupova podataka kemijskih spojeva ili proteina korištenih u modeliranju aktivnosti, svojstava ili strukturnih svojstava proteina (poput sekundarne strukture ili topologije membranskih proteina), molekularni deskriptori izvedeni iz kemijske strukture obično su korelirani s osnovnim svojstvima spojeva (npr. molekularna težina, veličina, oblik, broj specifičnih atoma, broj veza) ili proteina (npr. duljina slijedova, ukupan broj nekih specifičnih aminokiselina, postotak sekundarne strukture). Stoga, kako bi se mogla procijeniti stvarna razina nasumične točnosti (ili korelacije) modela, mora se osigurati da generirani slučajni podaci koji se koriste u simulacijama imaju strukturu i raspodjelu sličnu onima stvarnih ulaznih podataka.

Procjena razine nasumične točnosti klasifikacijskih problema s dva stanja (tj. dvije skupine svojstava/aktivnosti) bit će izložena u nastavku. U najjednostavnijem pristupu modeliranja sekundarne strukture membranskih proteina alfa vrste kao jedno stanje promatraju se dijelovi proteinskog lanca koji su u membrani (imaju sekundarnu strukturu α), a aminokiseline ostatka lanca svrstavaju se u nepravilnu sekundarnu strukturu i čine drugo stanje.

Sekundarna struktura α dijela proteina koji je u membrani organizirana je u jedan ili više segmenata od kojih se svaki sastoji od (uglavnom) $18-22$ susjedne aminokiseline. Ti se dijelovi nazivaju transmembranski segmenti, a aminokiseline koje ga čine označavaju se s M. Aminokiseline u dijelovima proteinskog lanca koji je smješten izvan membrane označavaju se s U. Pojednostavljena shema eksperimentalne sekundarne strukture membranskog proteina sa 100 aminokiselinskih ostataka i sjednim TM segmentom od 20 aminokiselina, dana je u tablici 13.

Tablica 13. Pojednostavljeni prikaz eksperimentalne sekundarne strukture membranskog proteina uj58_F, lanac F (dio lanca od 49 do 148 aminokiselina, ukupno 100 aminokiselina), i sekundarne strukture predviđene uravnoteženim modelom.[a]

Aminokiselinski lanac (ak lanac) Shema eksperimentalne sekundarne strukture (eksp. ss), M (membranski dio, α-uzvojnica)[b] Predviđena sekundarna struktura (predv. ss), M (membranski dio, α-uzvojnica)
1----------50
ak lanac: EALGNAEGISQEMSDATGIFVVRQLKGDRTGKIADIYQLNAQDASAMVLG
exp. ss: UUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUU
predv. ss: UUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUU
51----------100
ak lanac: TEFQLQPYDIVYVTTAPLVRWNRVISQLVPTISGVHDMTETVRYIKRWPN
exp. ss: UUUUUUUUUUUUUUUUUUUUUUUUUMMMMMMMMMMMMMMMMMMMMUUUUU
predv. ss: UUUUUUUUUUUUUUUUUUUUUMMMMMMMMMMMMMMMMMMMUUUUUUUUUU

[a]Duljina proteinskog slijeda: $N = 100$ aminokiselina, $N = n(\mathrm{U}) + n(\mathrm{M})$

[b]Ukupan broj aminokiselina eksperimentalno određen kao da je u stanju U (nepravilna sek. str.) i M (membranski dio, α-uzvojnica), uz $n(\mathrm{U}) = 80$ i $n(\mathrm{M}) = 20$.

Usporedbom stvarnih (eksperimentalnih) i predviđenih struktura, definiraju se parametri p, u, n i o:

- vrijednost parametra p (ispravno pozitivno predviđanje) definira se kao broj aminokiselina koje su u eksperimentalnoj sekundarnoj strukturi u membrani (M) i u sekundarnoj strukturi predviđenoj modelom također u membrani (M),
- vrijednost parametra u (pogrešno pozitivno predviđanje, engl. *underprediction*) broj je aminokiselina koji su u eskperimentalnoj sekundarnoj strukturi u membrani (M), a modelom su predviđeni izvan membrane (U),
- vrijednost parametra n (ispravno negativno predviđanje) broj je aminokiselina koji su u eskperimentalnoj sekundarnoj strukturi izvan membrane (U), a modelom su predviđeni izvan membrane (U),
- vrijednost parametra o (pogrešno negativno predviđanje, engl. *overprediction*) broj je aminokiselina koji su u eskperimentalnoj sekundarnoj strukturi izvan membrane (U), a modelom su predviđeni unutar membrane (M).

Na primjer, za lanac naveden iz tablici 13 vrijedi:

$p = 15$ – ispravno pozitivno predviđanje (eksp. M predviđeno kao M), podcrtano na eksperimentalnim i na predviđenim strukturama),

$u = 5$ – pogrešno pozitivno predviđanje (eksp. M predviđeno kao U),

$n = 75$ – negativno ispravno predviđanje (eksp. U predviđeno kao U),

$o = 5$ – pogrešno negativno predviđanje (eksp. U predviđeno kao M).

Očigledno je $p + u + n + o = N = 100$ aminokiselina, te $p + u = n(\mathrm{M}) = 20$ i $n + o = n(U) = 80$. U ovom slučaju kažemo da je predviđanje dobiveno modelom uravnoteženo, jer model predviđa isti broj stanja (klasa) M i U kao što je u eksperimentalnom slijedu.

Tablica 14. Tablica kontingencije za eksperimentalne i procijenjene (po modelu) sekundarne strukture membranskog proteina.

A) Elementi opće tablice kontingencije

	(proc/pred) M	(proc/pred) U	Σ redak (eksperimentalno)
(eksperimentalno) M	p	u	$p+u$
(eksperimentalno) U	o	n	$n+o$
Σ stupac (procjena ili predviđanje)	$p+o$	$n+u$	

B) Tablica kontingencije dobivena je iz eksperimentalnih i procijenjenih struktura membranskih proteina na temelju sheme strukture proteina iz tablice 13.

	(proc/pred) M	(proc/pred) U	Σ redak (eksperimentalno)
(eksperimentalno) M	15	5	*20*
(eksperimentalno) U	5	75	*80*
Σ stupac (procjena ili predviđanje)	20	80	

Kvaliteta predviđanja klasifikacijskoga modela s dva stanja može se opisati i tablicom kontigencije 2×2 (tablica 14), koja je temeljena na primjeru strukture proteina iz tablice 13. U idealnom slučaju (model koji ima predviđanje s točnošću od 100%), svaka aminokiselina koja je u eksperimentalnoj sekundarnoj strukturi u membrani modelom je predviđena u membrani. Nadalje, svaka aminokiselina izvan membrane modelom je i predviđena izvan membrane (nema pogrešnih predviđanja), pa vrijedi $u = 0,\ o = 0$ i $N = p + n$.

Uravnoteženi skup podataka i uravnoteženi modeli

Stvarni skupovi podataka s dvije klase obično imaju različite brojeve elemenata u svakoj od klasa. Međutim, u nekim je slučajevima moguće stvoriti idealno uravnotežen eksperimentalni skup podataka s istim brojem elemenata klase ($p + u = n + o$). Ako je to moguće postići, poželjno je koristiti takve podatke u razvoju i optimizaciji modela, jer se u tom slučaju obje klase ravnopravno tretiraju tijekom optimizacije modela. Drugi slučaj je uravnoteženi skup dobiven u procjeni (ili predviđanju), tj. kada se pomoću modela predviđa podjednak broj stanja u svakoj od dvije klase ($p + o = n + u$). Treći je slučaj model uravnotežen u predviđanju broja elemenata u svakoj od klasa. U tom slučaju imamo i $p + u = p + o$ (za stanje M) i $n + o = n + u$ (za stanje U), što daje $u = o$. Međutim, u i o ne moraju biti jednaki nuli i može biti $p + u \neq n + o$ (za eksperimentalni skup) ili $p + o \neq n + u$ (za predviđena stanja M i U). Dakle, dobro provedeno modeliranje završit će nakon što model dosegne ravnotežu između u i o u procjeni na skupu za optimizaciju modela. Samo u tom idealnom slučaju kada je $u = o$ moguće je dobiti maksimalnu točnost klasifikacije $Q_2 = 100\%$.

Polazeći od tablice kontigencije (tablica 14), moguće je definirati različite statističke parametre [66,69], a među njima je najjednostavniji ranije spomenuti parametar za procjenu točnosti modela Q_2:

$$Q_2 = \left(\frac{p+n}{p+n+u+o} \cdot 100\right)(\%) = \left(\frac{p+n}{N} \cdot 100\right)(\%)$$

gdje su: p, n, u i o definirani ranije u tablici 14, a $N = p + n + u + o$.

Stvarni ili nasumični model predviđa $(p + o)$ aminokiselina u stanju (klasi) M za $(p + u)$ eksperimentalno određenih aminokiselina u klasi M, i $(n + u) = N - (p + o)$ aminokiselina u klasi U za $(n + o)$ aminokiselina eksperimentalno određenih u klasi U. Tada je najvjerojatnija nasumična točnost dana izrazom:

$$Q_{2,rnd} = \left\{\left[\frac{p+u}{N}\cdot\frac{p+o}{N} + \frac{n+o}{N}\cdot\frac{n+u}{N}\right]\cdot 100\right\}(\%)$$

ili kraće

$$Q_{2,rnd} = \left\{\left[\frac{(p+u)\cdot(p+o)+(n+o)\cdot(n+u)}{N^2}\right]\cdot 100\right\}(\%)$$

Ovo je najvjerovatnija vrijednost parametra Q_2 koji se može dobiti proizvoljnim nasumičnim modelom.

Ako ekspermentalna sekundarna struktura jednog proteina (ili ukupni podaci za cijeli skup proteina) sadrži isti broj aminokiselina u stanjima M ili U, tada je $p+u = n+o = \frac{N}{2}$ pa se prethodni izraz može faktorizirati

$$Q_{2,rnd} = \left\{\left[\frac{(p+u)\cdot(p+o+n+u)}{N^2}\right]\cdot 100\right\}(\%) = \left(\frac{\frac{N}{2}\cdot N}{N^2}\cdot 100\right)(\%) = 50\%$$

To znači da je za podatke s podjednakim brojem M i U stanja u eksperimentalnoj strukturi najvjerojatnija vrijednost parametra $Q_{2,rnd} = 50\%$, i ta vrijednost ne ovisi o tome koliki je omjer brojeva stanja M i U u predviđenoj strukturi. Treba imati na umu da isto vrijedi i u slučaju kada je modelom predviđen podjednak broj stanja M i U ($p+o = n+u = \frac{N}{2}$), tada je ($Q_{2,rnd} = 50\%$), bez obzira na omjer brojeva aminokiselina u stanjima M i U u eksperimentalnoj strukturi.

Nadalje, ako se promatra uravnoteženi model, tj. model koji predviđa u sekundarnoj strukturi proteina broj aminokiselina u stanju M $(p+o)$ jednak onom u eksperimentalnoj strukturi $(p+u)$, iz čega proizlazi odgovarajuća jednakost (tj. $n+u = n+o$) i za stanje U, tada $Q_{2,rnd}$ postaje $Q_{2,rnd-bal}$:

$$Q_{2,rnd} = Q_{2,rnd-bal} = \left[\frac{(p+u)^2+(n+o)^2}{N^2}\cdot 100\right](\%)$$

Takav bi se model trebao dobiti nakon svakog ispravno provedenog postupka izbora i optimizacije modela. Ovom jednadžbom moguće je izračunati najvjerojatniju nasumičnu točnost za uravnoteženi (balansirani) model koji se planira razviti, i to samo na temelju broja aminokiselina u stanjima M i U u eksperimentalnoj strukturi, tj. $(p+u)$ za stanje M. Iz poznatog $(p+u)$, jednostavno se izračuna broj stanja U u eksperimentalnoj strukturi kao $n+o = N-p-u$. Također, iz ove jednadžbe slijedi da je za uravnoteženi model minimalna vrijednost $Q_{2,rnd} = 50\%$, kada su obje klase jednako zastupljene u eksperimentalnom skupu podataka $(p+u = n+o)$.

Ukoliko model nije uravnotežen pa predviđa veći broj aminokiselina u stanju M nego je taj broj u eksperimentalnoj strukturi, minimalna točnost može biti manja od 50% a maksimalna točnost prema parametru Q_2 uvijek je manja od 100% (rad u pripremi).

Konačno, razlika između stvarne točnosti modela Q_2 i njemu odgovarajuće nasumične točnosti $Q_{2,rnd}$ za uravnotežene ne modele računa se pomoću izraza:

$$\Delta Q_2 = Q_2 - Q_{2,rnd}$$

Ova vrijednost ima svoj maksimum $(\Delta Q_2)_{max} = 50\%$, za predviđanja uravnoteženim modelom kada je $u = o$ i $Q_{2,rnd} = 50\%$, i tada je jedino moguće postići maksimalnu točnost $Q_2 = 100\%$ koja se dobiva za $u = o = 0$.

Istodobno, uravnoteženi model razvijen na takvom eksperimentalnom skupu podataka najteži je problem za modeliranje, i analogan je problemu bacanja novčića koje se ponavlja N puta, pri čemu je $N = p+n+u+o$. Dakle, maksimalni raspon ΔQ_2 za razvoj i optimizaciju

modela (tj. ako pretpostavimo da je naš 'algoritam' nasumično pogađanje) od razine najvjerojatnije nasumične točnosti (50%) do razine maksimalno moguće stvarne točnosti (100%) iznosi 50%. Stvarni model za klasifikaciju dva stanja obično je razvijen na neuravnoteženom eksperimentalnom skupu stvarnih sekundarnih struktura s različitim brojem stanja M i U. Takvi modeli imat će (u pravilu) ΔQ_2 manji od 50%, što je ilustrirano na nekoliko primjera iz literature [46].

Problem koji se uočava kod parametara Q_2 i $Q_{2,rnd}$ njihova je neosjetljivost na razmjernu promjenu duljine proteinskog slijeda i broja aminokiselinskih ostataka u membrani, tj. ne prepozanaju ove promjene ako njihov omjer ostaje jednak. Na primjer, zamislimo da imamo proteinski lanac dug 100 aminokiselina koji ima jedan transmembranski segment duljine 20 aminokiselina, tada je:

$$Q_{2,rnd} = \frac{20^2 + 80^2}{100^2} = 0.68$$

Uzmimo drugi proteinski lanac koji ima dvostruko veću duljinu $N = 200$ i dva transmembranska segmenta svaki s po 20 aminokiselina (tj. ukupno 40 aminokiselina u stanju M). Parametar $Q_{2,rnd}$ za ovaj lanac jednak je:

$$Q_{2,rnd} = \frac{40^2 + 160^2}{200^2} = 0.68$$

Odavde vidimo da, iako imamo dvostruko dulji slijed s dvostruko više transmembranskih segmenata, parametar $Q_{2,rnd}$ ima istu vrijednost. Stoga, ovaj parametar nije osjetljiv na promjenu duljine proteinskog lanca (strukture) i nije u stanju ispravno razlikovati složenost struktura različitih duljina. Naime, intuitivno je jasno kako je znatno teže predvidjeti položaje dva TM segmenta na duljem slijedu nego jedan na dvostruko kraćem proteinskom slijedu. Kako bi se definirao parametar koji bi bio realističnija mjera složenosti, bilo bi korisno (ako je moguće) pronaći takav parametar koji će biti osjetljiv na promjenu duljine slijeda i broja TM segmenata, odnosno proporcionalne promjene duljine slijeda i broja aminokiselina u membrani.

Zaključci vezani uz parametar nasumične točnosti modela s dva stanja $Q_{2,rnd}$:

1) Razlika između parametara točnosti stvarnog modela Q_2 i odgovarajućeg slučajnog modela $Q_{2,rnd}$ korisni je parametar za provjeru kvalitete klasifikacijskih modela s dva stanja.
2) Parametar $Q_{2,rnd}$ neosjetljiv je na promjenu duljine slijeda proteinskog lanca, ako je omjer broja stanja u membrani i izvan nje nepromijenjen.
3) Postoji potreba za parametrom koji će moći razlikovati složenost struktura različitih duljina, na što je parametar $Q_{2,rnd}$ neosjetljiv.

3.1.2. Koncept nasumičnog modela, broj stanja i veza s entropijom

Kako bi se moglo procijeniti koji su reprezentativni skupovi membranskih proteina zahtjevniji (teži) u smislu pronalaženja modela za predviđanje položaja TM segmenata, osim navedenog parametra $Q_{2,rnd}$, potrebno je definirati koncepte za procjenu (i kvantificiranje) složenosti sekundarne strukture pojedinog proteinskog lanca. Složenost strukture iskazat će se preko broja mogućih ralizacija (W ili $\ln W$) modelne sekundarne strukture proteinskog lanca. U nastavku, najprije će se izložiti osnove statističke fizike koje su s tim u uskoj vezi.

3.1.2.1. Nasumični model na primjeru mnoštva čestica u statističkoj fizici

Termodinamički procesi u sustavima s mnoštvom čestica (npr. plinu) mogu se objasniti zakonima klasične fizike koristeći Hamiltonove jednadžbe gibanja za svaku česticu (ili molekulu plina) [70]

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \dot{p}_i = -\frac{\partial H}{\partial q_i}.$$

Da bi se riješile Hamiltonove jednadžbe gibanja trebalo bi poznavati početne uvjete koordinata i impulsa $q_i(t_0)$ i $p_i(t_0)$, čime bi se u cijelosti odredile njihove vrijednosti u vremenu $t > t_0$. No, kako je u mehanici moguće koristiti inverziju vremena, postavlja se pitanje kako onda objasniti termodnamičke nepovratne procese. U termodinamici se mjere makroskopske veličine sustava, a ne mikrostanje pojedine čestice koje grade taj sustav, pa kažemo da svako makrostanje sustava može biti realizirano preko velikog broja različitih mikrostanja. Kako bi se odredilo u koje će makrostanje sustav prijeći, mora se zapravo odrediti koje su statističke težine pojedinih mikrostanja. Pritom, sustav će težiti k onim makrostanjima koja su statistički najvjerojatnija. Na ovaj način definira se termodinamička vjerojatnost preko broja različitih mikrostanja koja ostvaruju dano makrostanje, a ta vjerojatnost odgovara statističkoj težini stanja. Makrostanja iste statističke težine mogu se sastojati od različith kombinacija mikrostanja.

Analizirajmo primjer u kojem se pitamo na koliko se različitih načina može ostvariti razmještanje N_i čestica u g_i prostornih kutija. Pretpostavimo da su čestice istih fizikalnih svojstava, a prostor u kojemu se nalaze čestice promatramo kao sustav. Podijelimo čestice na skupine po određenoj fizikalnoj karakteristici (npr. ista energija). Neka je broj čestica s energijom E_i jednak N_i $(i = 1, 2, 3, \cdots)$, gdje vrijedi:

$$N = \sum_i N_i$$

Standardnim modelom matematičke statistike najprije analizirajmo slučaj kada nema degeneracije po energijama, i pritom se promatra samo razmještanje čestica u prostoru.

Broj različitih načina izbora jedne čestice od njih N je $N!$. Ako biramo dvije čestice od početnih N, tada će broj načina odabira biti jednak N za prvu i od preostalih $(N-1)$ za drugu, tj. ukupno $N \cdot (N-1)$. Ako se odabire n čestica, tada je broj svih mogućih načina izbora jednak

$$N \cdot (N-1) \cdot (N-2) \cdots (N-n+1) = \frac{N!}{(N-n)!}$$

Ako pri izboru tih n čestica nije bitan redoslijed, tada prethodni izraz treba podijeliti s ukupnim brojem permutacija n čestica što je jednako $n!$. To za ukupni broj mogućih izbora daje

$$\frac{N!}{(N-n)! \cdot n!} = \binom{N}{n}$$

što je zapravo jednako binomnom koeficijentu kojim se računa broj svih kombinacija (realizacija) na koje se može uzeti n elemenata iz početnih N.

Pretpostavimo sada da se izabrane čestice smještaju u prostoru koji ima k kutija, i to tako da N_1 čestica smještamo u ćeliju *1*, N_2 čestica smještamo u ćeliju *2*, i tako dalje dok sve izabrane čestice ne smjestimo u ćelije. Primjenom prethodnog izraza, izračunava se broj mogućih načina na koji se može izabrati prvih N_1 čestica od ukupno N čestica:

$$W(N_1) = \frac{N!}{(N-N_1)! \cdot N_1!} = \binom{N}{N_1}$$

Analogno dobivamo izraz za broj mogućih načina na koji se može izabrati narednih N_2 čestica koje biramo od preostalih $(N - N_1)$ čestica:

$$W = \frac{N!}{(N-N_1)! \cdot N_1!} \cdot \frac{(N-N_1)!}{(N-N_1-N_2)! \cdot N_2!} \cdot \ldots \cdot \frac{(N-N_1-\cdots-N_{k-1})!}{(N-N_1-\cdots-N_k)! \cdot N_k!}$$

Nakon skraćivanja istih faktorijela, konačno se dobiva

$$W = \frac{N!}{N_1! \cdot N_2! \cdot \cdots \cdot N_k! \cdot (N-N_1-\cdots-N_k)!}$$

Ako sve čestice rasporedimo u ćelije, tada je broj preostalih čestica jednak nuli, tj.

$$(N - N_1 - \cdots - N_k) = 0$$

te se prethodni izraz (zbog $0! = 1$) mijenja u

$$W = \frac{N!}{N_1! \cdot N_2! \cdot \cdots \cdot N_k!} = N! \cdot \prod_{i=1}^{k} \frac{1}{N_i!}$$

Ovo je izraz kojim se računa broj mogućih načina razmještanja *n* čestica u *k* kutija. Pritom, zanemaruju se moguće permutacije smještanja čestica u različitim kutijama.

Pogledajmo sada kako se mijenja broj načina razmještanja čestica ako je svaka pojedina kutija degenerirana. Degeneracija se može predočiti na način da zamislimo kako svaka kutija u kojoj je N_i čestica ima g_i podkutija (odnosno onoliko podkutija kolika je degeneracija). Tada se za svaku ćeliju dobije još i broj mogućih razmještaja N_i čestica na g_i podkutija, što je jednako ${g_i}^{N_i}$. Ovo razmatranje vodi na Boltzmannovo prebrojavanje [71] koje je dano izrazom::

$$B = \frac{N!}{N_1! \cdot N_2! \cdot N_3! \cdots} \cdot ({g_1}^{N_1}) \cdot ({g_2}^{N_2}) \cdot ({g_3}^{N_3}) \cdots = N! \cdot \prod_i \frac{{g_i}^{N_i}}{N_i!}$$

Veličinu B nazivamo termodinamičkom vjerojatnošću. Ukupni broj realizacija neke raspodjele dobivamo tako što broj permutacija svih čestica dijelimo umnoškom broja permutacija čestica na pojedinoj razini s istom fizikalnom karakteristikom (npr. energijom). Pritom, svaku od N_i čestica na istoj razini u bilo kojoj prostornoj ćeliji možemo razmjestiti na ${g_i}^{N_i}$ načina.

Svakako, ovaj se problem može razmatrati i tako da se definira opći oblik vjerojatnosti (p_i) realizacije nekog stanja *i*. Ako se određeni sustav može naći u mnoštvu stanja, tada se računa vjerojatnost nalaženja sustava u pojedinom stanju kao omjer broja realizacija tih (pojedinih) stanja u odnosu na ukupni broj svih mogućih stanja. Koristeći ovako definiranu vjerojatnost [70,71], Ludwig Boltzmann i J. Willard Gibbs definirali su u teoriji statističke mehanike entropiju ovim izrazom

$$S = -k_B \sum_i p_i ln(p_i)$$

gdje je p_i vjerojatnost mikrostanja *i* uzetog iz ravnotežnog ansambla. S druge pak strane, u informacijskoj teoriji Claude E. Shannon definirao je etropiju izrazom

$$H = -\sum_i p_i log_b(p_i)$$

gdje je p_i vjerojatnost nalaženja informacije m_i unutar informacijskog prostora *M*, a *b* je baza logaritma. Uobičajeno se za bazu logaritma koriste vrijednosti 2 (za Shannonovu entropiju ili bit), zatim baza prirodnog logaritma *e*, te broj 10 (Hartleyeva entropija).

Ako su sva stanja jednako vjerojatna (mikrokanonski ansambl), tada se izraz za statističku termodinamičku entropiju uvođenjem izraza $p_i = 1/W$ (gdje je W broj mikrostanja), pojednostavljuje na

$$S = -k_B \sum_{i=1}^{W} p_i ln(p_i) = -k_B \sum_{i=1}^{W} \frac{1}{W} ln\left(\frac{1}{W}\right) = -k_B \cdot \left(\frac{1}{W}\right) \cdot W \cdot (-lnW) = k_B lnW$$

To je poznati izraz iz statističke fizike koji povezuje entropiju s brojem stanja sustava. Entropija sustava to je veća što je veći broj mogućih realizacija stanja sustava.

3.1.2.2. Nasumični model na primjeru sekundarne strukture proteina

Ako promatramo stvarne biološke makromolekule kao što su proteini (DNA, RNA ili druge molekule) u opisu njhove strukture primjenjuje se koncept broja mogućih strukturnih konformacija (stanja). Taj je koncept u vezi s konformacijskom entropijom strukture. Za računanje broja konformacija u trodimenzionalnoj strukturi proteina potrebno je promatrati dihedralne kutove na Cα atomu svake aminokiseline (koji definiraju smjer glavnog lanca). Osim ove konformacijske entropije strukture moguće je definirati i konformacijsku entropiju jedne strukturne konformacije (realizacije), npr. na primjeru sekundarne strukture proteina. Za jednu odabranu sekundarnu strukturu proteinskog lanca možemo se pitati: koliki je broj mogućih realizacija takve modelne strukture, pri čemu dopuštamo razmještanje elemenata sekundarne strukture na sva moguća mjesta u proteinskom lancu. Razmještanje se provodi sve dok se ne iscrpe sve moguće realizacije. Ako se promatra sekundarna struktura proteina sa samo dva stanja, tada konformacijska entropija (broj realizacija modelne strukture) u prvoj aproksimaciji ovisi o: (1) zastupljenosti pravilne sekundarne strukture α u ukupnoj strukturi i (2) duljini proteinskog lanca.Taj će se problem analizirati u nastavku u analogiji s prikazanim primjerom iz statističke fizike (3.1.2.1), a konačni je cilj izvesti izraz koji procjenjuje složenost (kompleksnost, zahtjevnost) sekundarne strukture proteina.

Kako su proteini građeni od osnovnih elemenata (aminokiselina), može se (u prvoj aproksimaciji) uzeti u obzir da je svaka aminokiselina jedna čestica u modelu. Nadalje, svaka se aminokiselina može ponavljati proizvoljni broj puta u definiranom 'prostoru' koji je jednak duljini proteinskog slijeda. Na ovaj se način model raspodjele čestica u kutiji iz statističke fizike može primijeniti u analizi i kvantificiranju složenosti sekundarne strukture proteina.

Razmotrimo sada pojednostavljeni model strukture integralnih membranskih proteina, u kojem se zapravo neće gledati vrsta aminokiseline u proteinskom lancu, nego položaj aminokiseline u odnosu na membranu. Za integralne membranske proteine α-vrste, kod kojih aminokiseline u dijelu lanca koji odgovara TM segmentu poprimaju sekundarnu strukturu α a one izvan membrane poprimaju nepravilnu sekundarnu strukturu, definiramo pojednostavljeni model s dva stanja ('u membrani' ili 'izvan membrane').

Ukoliko neka struktura ima veliki broj mogućih realizacija, možemo reći da je ona zahtjevnija (teža) za razumjeti (tj. teže je naučiti pravila strukturiranja koja su u podlozi takve strukture). U postupku učenja, algoritmom se nastoji pronaći optimalni model koji će, na temelju specifičnosti (1) aminokiselinskog slijeda, (2) svojstava aminokiselina te (3) evolucijske informacije za određeni protein, biti u stanju što je moguće bolje 'naučiti' ili 'reproducirati' njegovu stvarnu sekundarnu strukturu. Algoritam koji u postupku učenja pronalazi optimalni model, teže će pronaći pravila koja su potrebna za predvidjeti strukturu ako je ona složenija, pa će optimizacija modela biti zahtjevnija. Zamislimo li novi proteinski lanac za koji želimo predvidjeti sekundarnu strukturu s pomoću modela razvijenoga ranije na drugom skupu sekundarnih struktura proteina. Za očekivati je da će složeniju strukturu model predviđati s manjom točnošću.

3.1.3. Binomni nasumični model i procjena složenosti sekundarne strukture proteina

Proteinski slijed sastavljen je od niza aminokiselina koje su vezane jedna na drugu definiranim redoslijedom zadanim genom koji kodira protein. Svaka aminokiselina može se vezati (može imati u susjedstvu, s lijeve ili desne strane) na istovrsnu aminokiselinu ili na bilo koju od preostalih 19 aminokiselina. Međutim, redoslijedi poznatih prirodnih proteinskih slijedova nisu sasvim nasumični redoslijedi nego samo oni koji daju, za živi svijet kakav poznajemo, funkcionalne 3D strukture proteina. Nadalje, različita je učestalost:

(1) pojavljivanja svake od 20 aminokiselina u poznatim prirodnim i funkcionalnim proteinima, i
(2) pojavljivanja parova susjednih aminokiselina.

Pripomenimo ovdje kako postoji $20 \cdot 20 - 20 = 380$ mogućih parova susjednih aminokiselina, i to stoga što proteinski lanac nije simetričan s obzirom na N- i C-kraj pa tako npr. par susjednih aminokiselina alanin–glicin nije jednak kao i par glicin–alanin. Jedino se ne može razlikovati 20 parova identičnih susjednih aminokiselina.

Za proteinski lanac može se izračunati ukupan broj mogućih realizacija (modelnih konformacija) sekundarne strukture (u prvoj aproksimaciji) na način da svaku aminokiselinu u proteinskom lancu promatramo kao jednu nezavisnu česticu. Tako se lanac koji ima ukupno M aminokiselina u membranskom dijelu promatra kao sustav od M čestica i na njega se primjenjuje statistički model nasumične raspodjele čestica u kutiji. Tada je raspoloživi prostor jednak duljini proteinskog slijeda N (N kutija), i u tom prostoru raspoređujemo aminokiseline kao čestice. Ako je pojedina aminokiselina proteinskog lanca unutar membranskog dijela, kažemo da je čestica u kutiji. U suprotnom, ako je aminokiselina izvan membranskog dijela, kažemo da je kutija prazna.

Uz pretpostavku da u N kutija raspoređujemo M čestica koje se ne razlikuju, jer se razmatra samo položaj čestice u odnosu na membranu (to bi značilo da u proteinskom slijedu duljine N imamo M aminokiselina u membrani), broj mogućih realizacija definira se binomnom raspodjelom:

$$W(N, M) = \binom{N}{M}.$$

U modelu nasumične raspodjele proizvoljnog broja čestica u N kutija uz uvjet da je moguće staviti samo jednu česticu u jednu kutiju (tj. da nema degeneracije), ukupan broj realizacija raspodjela čestica jednak je zbroju svih binomnih članova u rasponu od 0 do N [72]:

$$W = \sum_{M=0}^{N} \binom{N}{M} = 2^N$$

Normirane vrijednosti dobiju se kada se broj mogućih realizacija rasporeda određenog broja čestica M podijeli s ukupnim brojem svih mogućih konformacija. Normirana vrijednost jednaka je vjerojatnosti nalaženja pojedine raspodjele M čestica, a definira se izrazom:

$$P(N, M) = \frac{1}{2^N} \cdot \binom{N}{M}.$$

Binomna raspodjela može se primijeniti na sekundarnu strukturu proteina na način da se aminokiseline smještene u membranu smatraju česticama prve vrste (tj. $n_1 = M$), dok se aminokiseline koje nisu smještene u membrani smatraju česticama druge vrste (njih je $n_2 = N - M = N - n_1$), gdje je N duljina proteinskog slijeda. Na ovako definiranom modelu, kojeg nazivamo binomni nasumični model, razmatramo na koliko se načina mogu preraspodijeliti aminokiseline u membranu (odnosno izvan nje) i to nazivamo i brojem mogućih realizacija modelne strukture prema binomonom nasumičnom modelu. Potpuno je svejedno hoćemo li

razmatrati razmještaj aminokiselina u membrani (binomni izraz za $n_1 = M$), ili za preraspodjelu aminokiselina izvan membrane (binomni izraz za $n_2 = N - n_1$) jer vrijedi

$$\boxed{W(N, n_1) = \binom{N}{n_1} = \frac{N!}{n_1! \cdot (N - n_1)!} = \frac{N!}{n_1! \cdot n_2!}}$$

$$W(N, n_2) = \binom{N}{n_2} = \frac{N!}{n_2! \cdot (N - n_2)!} = \frac{N!}{n_2! \cdot n_1!}$$

Jednakost između $W(N, n_1)$ i $W(N, n_2)$ slijedi iz simetričnosti binomnih koeficijenata, odnosno iz izraza:

$$\binom{N}{k} = \binom{N}{N - k}$$

Izraz za $W(N, M)$ daje broj realizacija modelne strukture membranskog proteina prema binomnom nasumičnom modelu, i korišten je u disertaciji na modelnim strukturama stvarnih proteina. Vjerojatnost stanja kada je *M* aminokiselina u membrani tada je:

$$P(N, n_1) = P(N, M) = \frac{1}{2^N} \cdot \frac{N!}{n_1! \cdot n_2!} = \frac{1}{2^N} \cdot \frac{N!}{M! \cdot (N - M)!}$$

Ovaj izraz zapravo znači da je vjerojatnost jednaka kada protein s *N* aminokiselina ima njih *M* u membrani ili $(N - M)$ aminokiselina u membrani.

U slučaju razlikovanja čestica, dobile bi se dvije različite raspodjele koje su međusobno simetrične:

$$W(N, n_1) = \binom{N}{n_1} \cdot n_1! = \frac{N!}{n_1! \cdot (N - n_1)!} \cdot n_1! = \frac{N!}{n_1! \cdot n_2!} \cdot n_1! = \frac{N!}{n_2!}$$

$$W(N, n_2) = \binom{N}{n_2} \cdot n_2! = \frac{N!}{n_2! \cdot (N - n_2)!} \cdot n_2! = \frac{N!}{n_2! \cdot n_1!} \cdot n_2! = \frac{N!}{n_1!}$$

I te raspodjele imaju jednak broj članova. Za velike *N* ($N \gg 0$): zbroj svih članova raspodjele (s razlikovanjem čestica) jednak je

$$\sum_{k=0}^{N} W(N, k) = \sum_{k=0}^{N} \binom{N}{k} \cdot k! = \sum_{k=0}^{N} \frac{N!}{(N - k)!} = N! \cdot \sum_{k=0}^{N} \frac{1}{k!} \cong N! \cdot e.$$

Vjerojatnost jednog stanja (tj. stanja n_1 odnosno 'aminokiselina je u membrani') bila bi:

$$P(N, n_1) = \frac{W(N, n_1)}{\sum_{k=0}^{N} \binom{N}{k} \cdot k!} = \frac{\frac{N!}{n_2!}}{N! \cdot e} = \frac{1}{(N - n_1)! \cdot e} = \frac{1}{n_2! \cdot e}.$$

Ovaj izraz daje maksimalnu vjerojatnost nalaženja aminokiselina u membrani kada je $n_2 = 0$, odnosno kada vrijedi $n_1 = M = N$.

3.1.4. Segmentni nasumični model i procjena složenosti sekundarne strukture proteina

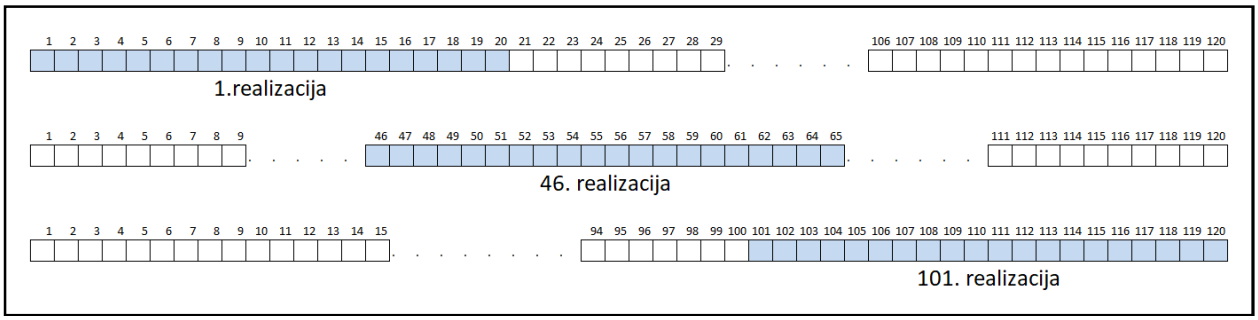
Proteinski slijed u sebi sadrži dijelove koji imaju uređenu strukturu (npr. alfa uzvojnicu ugrađenu u membranu) koja se na osnovnoj razini naziva sekundarna struktura. Uređeni dijelovi sekundarne strukture smatraju se relativno nepromijenjeni u odnosu na ostatak (tj. nepravilni dio)

strukture i kompleksnost se računa na način da se ukupan broj modelnih konfiguracija smanjuje za moguće modelne konfiguracije u kojima bi se dijelovi sa uređenom sekundarnom strukturom međusobno zamijenjivali. Drugim riječima, te segmente uređene sekundarne strukture držimo nepromijenjenima (fiksnima) kao jednu česticu (element) u modelu.

3.1.4.1. Izvod izraza za prebrojavanje realizacija u sekundarnoj strukturi (segmentni nasumični model).

Postavlja se pitanje kako se mijenja broj nasumičnih modelnih konformacija strukture ako se u ukupnoj duljini slijeda lanca uzme segment duljine d u kojemu više nema permutacija aminokiselina nego se on smatra jednim nepromjenjenim dijelom (segmentom)? U toj se slici segment ponaša kao jedna jedina čestica, a prostor koji ostaje na raspolaganju za dobivanje novih konformacija reducira se za duljinu tog segmenta umanjenu za jedan $(d-1)$.

Npr. za proteinski slijed duljine 120 aminokiselina ($N = 120$), u kojem imamo jedan segment ($s = 1$) duljine 20 aminokiselina ($d = 20$) dobiva se 101 moguća realizacija položaja takvoga segmenta. Ovaj broj realizacija možemo dobiti tako što zbroju duljine slijeda i broja segmenata oduzmemo duljinu segmenata, što za jedan segment daje $100 + 1 - 20 = 101$. Moguće realizacije za ovaj slučaj prikazane su na slici 13.



Slika 13. Grafički prikaz realizacija u segmentnom nasumičnom modelu za slijed duljine 120 aminokiselina s jednim segmentom duljine 20 aminokiselina

Broj realizacija za slijed s jednim segmentom mogao bi se napisati preko binomnog obrasca u ovisnosti o brojevima koji karakteriziraju lanac duljine N s jednim ($s = 1$) segmentom duljine d pomoću izraza:

$$W(N, d, s = 1) = \binom{N + 1 - d}{1},$$

što u slučaju duljine proteinskog slijeda $N = 120$, s jednim segmentom ($s = 1$) duljine 20 aminokiselina ($d = 20$) iznosi:

$$W(120, 20, 1) = \binom{120 + 1 - 20}{1} = 101.$$

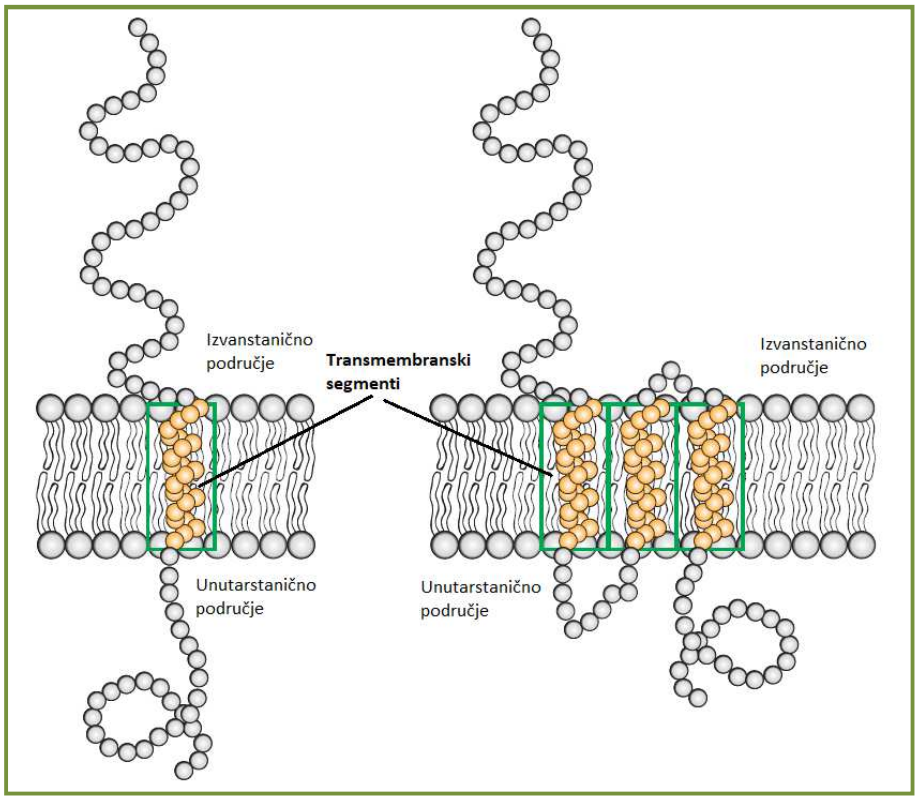
Ako bi se isti problem razmatrao pomoću jednostavnog modela proizvoljne raspodjele 20 čestica (koje se ne razlikuju) u prostoru duljine 120, broj svih mogućih konformacija modelne strukture bio bi neusporedivo veći:

$$W(120, 20) = \binom{120}{20} \cong 2.95 \cdot 10^{22}.$$

Proizlazi da se broj svih nasumičnih realizacija modelne srukture koji se mogu dobiti raspoređivanjem jednog segmenta (duljine 20) u lancu duljine $N = 120$, smanjio za 19 redova veličina u odnosu na izračun prema binomoj raspodjeli (nasumični model, ili ne-segmentni

nasumični model). To je potpuno razumljivo, jer informacija o tome da imamo jedan segment s česticama koje ne mogu mijenjati svoj položaj unutar određenog segmenta duljine d nužno, zbog te dodatne uređenosti, vodi na smanjenje broja mogućih modelnih konformacija.

Pogledajmo sada što se događa ako u slijedu imamo više od jednog segmenta. Na slici 14 shematski su prikazani proteinski lanci s jednim i tri TM segmenta. Zanima nas mogu li se brojevi realizacija modelnih struktura ova dva proteinska lanca iskazati zajedničkim izrazom.



Slika 14. Shematski prikaz proteina s jednim i tri transmembranska segmenta.

Ako se uzme u obzir da je u proteinskom lancu duljine N raspoređeno s segmenata od kojih je svaki jednake duljine d, postavlja se pitanje može li se gornji izraz za jedan segment poopćiti i na s segmenata (uz pretpostavku da segmente istih duljina ne razlikujemo). Tada bi izraz za jedan segment poprimio sljedeći oblik:

$$W(N,d,s) = \binom{N+s-d\cdot s}{s} = \binom{N-(d-1)\cdot s}{s}.$$

Uz pretpostavku da su duljine proteinskih lanaca sa slike 14 jednake (npr. $N = 300$) i da su duljine segmenta jednake (npr. $d = 24$), može se vidjeti kolika je razlika u brojevima mogućih preraspodjela segmenata u tim lancima u odnosu na ne-segmentni nasumični model. Za prvi lanac s jednim segmentom broj modelnih konformacija jednak je

$$W(300,24,1) = \binom{300+1-(24\cdot 1)}{1} = \binom{277}{1} = 277.$$

Postupak bismo nastavili za drugi proteinski segment u lancu uzimajući da prvi segment zauzima stalno mjesto i gledajući koliko mogućih realizacija imamo za pomake drugog segmenta. Pritom, prvi i drugi segment mogu biti jedan do drugoga bez ijednog mjesta razmaka. Prvi segment zauzima 24 mjesta, pa drugi segment možemo smjestiti na $300 - 24 = 276$ preostalih slobodnih mjesta. Kako drugi segment također zauzima 24 mjesta, za pomake drugog segmenta ukupno preostaje $n = 300 - 24 - (24 - 1) = 253$ mjesta. Potom pomaknemo prvi segment za jedno

mjesto, pa imamo dodatne 252 realizacije položaja drugog segmenta. Postupak nastavimo tako da prvi segment pomičemo po jedno mjesto sve dok možemo vršiti pomake. Nakon što smo izvršili pomake prvog segmenta za 24 mjesta, drugi segment može smjestiti na položajima i prije prvoga. Međutim, te realizacije ne brojimo dodatno jer, po početnoj pretpostavci, ne razlikujemo segmente istih duljina. Na ovaj način dobije se ukupan broj realizacija kao zbroj svih brojeva od 1 do 253:

$$\sum_{k=1}^{253} k = 32\,131.$$

Zbroj prvih n prirodnih brojeva dan je izrazom

$$\sum_{k=1}^{n} k = \frac{n \cdot (n+1)}{2},$$

koji možemo prikazati i preko binomnog obrasca kao:

$$\binom{n+1}{2} = \frac{(n+1)!}{2! \cdot (n+1-2)!} = \frac{(n+1) \cdot n}{2}.$$

Kako je broj n jednak

$$n = 300 - 24 - (24 - 1) = 253,$$

to je

$$n + 1 = 300 - 24 - (24 - 1) + 1 = 300 + 2 - 24 \cdot 2 = 254.$$

Na ovaj način broj realizacija u našem slučaju za dva segmenta može se napisati kao:

$$W(300, 24, 2) = \binom{300 + 2 - (24 \cdot 2)}{2} = \binom{254}{2} = 32\,131,$$

odnosno općenito

$$W(N, d, s) = \binom{N + s - d \cdot s}{s}.$$

U slučaju da razlikujemo prvi i drugi segment, broj mogućih realizacija povećao bi se za sve moguće zamjene segmenata (za s segmenta to je $s!$). Poopćenje izraza za slučaj razlikovanja segmenata dao bi:

$$W(N, d, s) = \binom{N + s - d \cdot s}{s} \cdot s!,$$

$$W(300, 24, 2) = \binom{300 + 2 - (24 \cdot 2)}{2} \cdot 2! = \binom{254}{2} \cdot 2! = 64\,262.$$

Poopćenjem na slučaj proteinskog lanca (npr. sa slike 14.) s tri segmenta (koja model ne razlikuje), dobiva se sljedeći broj modelnih konformacija strukture

$$W(300, 24, 3) = \binom{300 + 3 - (24 \cdot 3)}{3} = \binom{231}{3} = 2\,027\,795.$$

U slučaju razlikovanja segmenata taj broj iznosi

$$W(300, 24, 3) = \binom{300 + 3 - (24 \cdot 3)}{3} \cdot 3! = \binom{231}{3} \cdot 3! = 12\,166\,770.$$

Proizilazi da broj mogućih modelnih konformacija izrazito raste s brojem TM segmenata. Porastom broja TM segmenata (porast veličine s u binomnom članu) proporcionalno raste i broj članova u umnošku razvoja binomnog člana. S druge strane, porastom duljine slijeda (porast N) rastu samo članovi u umnošku, dok broj članova u umnošku ostaje nepromijenjen. Zbog toga, promjena broja mogućih modelnih konformacija (realizacija) modelne strukture ima znantno brži porast pri povećanju broja TM segmenata uz konstantni *N*, nego porastom *N* uz konstantni broj TM segmenata *s*.

Prethodno razmatranje uzima u obzir samo segmente jednakih duljina. U narednim razmatranjima izvršiti će se poopćenja na proizvoljne duljine segmenata i proizvoljan broj segmenata. Pritom će se pretpostavljati da broj segmenata (teorijski) može rasti sve dok je zbroj svih njihovih duljina manji od duljine proteinskoga slijeda.

3.1.4.2. Segmentni nasumični model za proizvoljni broj segmenata

Pretpostavimo da u slijedu imamo *s* segmenata, tada je ukupna duljina slijeda koja se nalazi u segmentima (uz oznaku d_j za duljinu j-tog segmenta) jednaka

$$d(seg) = \sum_{j=1}^{s} d_j$$

Postavlja se pitanje – koliko će se promijeniti broj konformacija zbog ovog ograničenja? Pritom pretpostavljamo da se pojedini segmenti ne smiju preklapati, a mogu biti jedan do drugoga, bez razmaka. Također, potrebno je uzeti u obzir da se ne može razlikovati segmente istih duljina. Tada je broj različitih realizacija definiran izrazom:

$$W(N,s) = \binom{N + s - d(seg)}{s} \cdot \frac{s!}{PFC},$$

gdje faktor PFC u drugom dijelu izraza dolazi usljed smanjenja broja mogućih realizacija strukture zbog nemogućnosti razlikovanja istovrsnih segmenata. Taj faktor zapravo je umnožak faktorijela broja pojedinih klasa s TM segmentima istih duljina ili skraćeno *PFC* i definiran je kao:

$$PFC = \prod_{k} klasa(\mathrm{TM}_k)!$$

Klasa TM segmenata istih duljina definira se kao broj TM segmenata istih duljina u proteinskom slijedu $[klasa(\mathrm{TM}_k)]$. Na primjer, ukoliko protein ima 3 TM segmenta duljine 19, dva TM duljine 20 i tri TM segmenta duljine 21, onda je faktor $PFC = 3! \cdot 2! \cdot 3! = 72$.

U slučaju razlikovanja segmenata istih duljina, broj konformacija povećao bi se jer bi vrijednost faktora *PFC* (koji je produkt faktorijela klasa) u tom slučaju bio

$$PFC = 1! \cdot 1! \cdot 1! = 1$$

i prethodni izraz poprimio bi vrijednost

$$W(N,s) = \binom{N + s - d(seg)}{s} \cdot s!.$$

Razmotrimo primjer slijeda u kojem su svi segmenti jednakih duljina i koji se međusobno ne razlikuju. Tada bi koeficijent PFC bio jednak $s!$, a duljina svih segmenata bila bi jednaka $d(seg) = s \cdot d$. Tada bi konačni izraz za broj mogućih realizacija nasumičnog smještanja segmenata u proteinskom slijedu bio

$$W(N,d,s) = \binom{N+s-d(seg)}{s} \cdot \frac{s!}{PFC} = \binom{N+s-d\cdot s}{s} \cdot \frac{s!}{s!} = \binom{N-(d-1)\cdot s}{s},$$

što je zapravo već izvedeni izraz za pojednostavljeni slučaj segmenata jednakih duljina.

3.1.4.3. Segmentni nasumični model s razmacima između segmenata

Razmatrajući razmake između TM segmenata u eksperimentalnim strukturama svih integralnih membranskih proteina α vrste uočeno je kako se jako rijetko jedan TM segment nastavlja izravno nakon drugoga. Zapravo je uobičajeno da se između dva transmembranska segmenta pojavljuje nekoliko (tri ili četiri) aminokiseline, često u strukturi zavoja. U slučaju da se želi dobiti izraz u kojemu se isključuju ona stanja s položajima TM segmenata međusobno udaljenim manje od *r* mjesta, tada je broj slobodnih mjesta za pomak između svaka dva TM segmenta umanjen za *r*. Zapravo, možemo pojednostavljeno reći kako se segmenti u tom slučaju ponašaju kao da imaju duljinu uvećanu za razmak *r* (izuzev jednoga segmenta). Ukupan broj zauzetih mjesta tada je

$$r \cdot (s-1),$$

a gornji izraz za model u kojemu se ne razlikuju TM segmenti jednakih duljina poprima opći oblik:

$$W[N, d(seg), s, PFC, r] = \binom{N+s-d(seg)-r\cdot(s-1)}{s} \cdot \frac{s!}{PFC}.$$

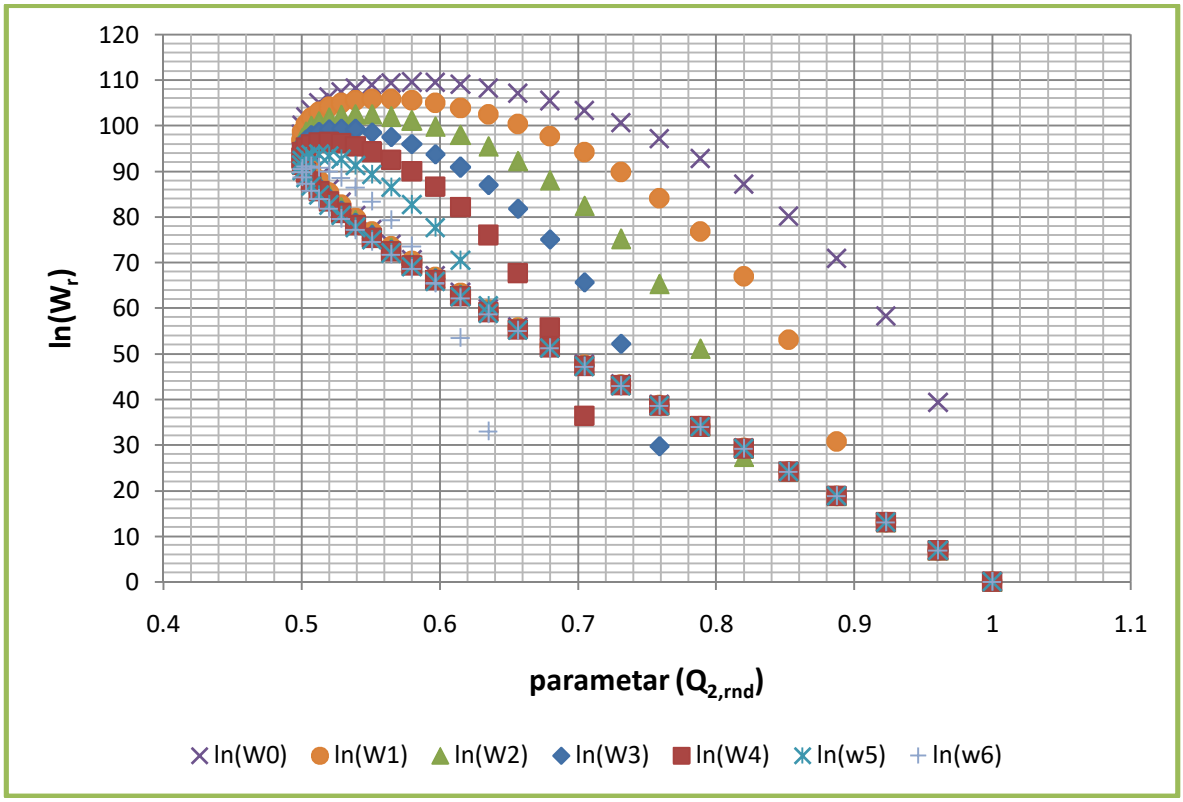
Zaključak: Ukupni broj kombinacija po ovome modelu ovisi o:

- ukupnoj duljini proteinskog slijeda (*N*)
- broju TM segmenata (*s*),
- zbroju njihovih duljina *d*(*seg*),
- produktima faktorijela koje čine brojevi istovrsnih TM segmenata u jednoj klasi (*PFC*), gdje klasa predstavlja npr. ukupni broj TM segmenata jednake duljine i
- razmaku između TM sgmenata po modelu (*r*)

Važno je napomenuti da konačni izraz ne ovisi posebno o duljinama pojedinih segmenata $[s(i)]$, nego samo o zbroju njihovih duljina *d*(*seg*).

Nadalje, ako razlikujemo sve segmente (a moramo ih razlikovati ako su različitih duljina), tada je ovo i konačni broj kombinacija (razmještanja), odnosno ukupni broj realizacija modelne strukture. U slučaju segmenata jednakih duljina, ukupan broj realizacija modelne strukture bio bi umanjen za broj mogućih kombinacija zbog međusobnih zamjena segmenata jednakih duljina, a on se iskazuje faktorijelom broja TM segmenata unutar svake klase segmenata jednakih duljina. Ako nema niti jedan par TM segmenata istih duljina, ili ako se razlikuju svi segmenti, u konačnom je izrazu $PFC = 1$.

Na slici 15. prikazane su vrijednosti logaritma (po prirodnoj bazi) broja mogućih realizacija W_r (gdje *r* poprima vrijednosti od 1 do 6) u ovisnosti u parametru $Q_{2,rnd}$, za duljinu slijeda $N = 1000$ i duljinu TM segmenata $d = 20$, za segmentni nasumični model s definiranim razmacima između segmenata uz r. Uočava se pad broja mogućih realizacija u segmentnom nasumičnom modelu s porastom minimalnog razmaka *r* između TM segmenata. To je potpuno razumljivo, jer nemogućnosti raspoređivanja segmenata na položajima minimalnih razmaka smanjuje ukupni 'prostor' (tj. broj mjesta u slijedu) za pomake segmenata, a time i broj mogućih realizacija.



Slika 15. Logaritam broja mogućih realizacija $\ln(W_r)$ za segmentni nasumični model s definiranim minimalnim razmacima r od 0 do 6 u ovisnosti u parametru $Q_{2,rnd}$ za duljinu slijeda $N = 1000$ i duljinu TM segmenata $d = 20$.

Umjesto TM segmenata, ovakav se model može i generalizirati na način da se svaki uređeni dio sekundarne strukture proteinskog slijeda (segment) može promatrati očuvanim dijelom. Tada bi se sličan izraz mogao primjeniti i na općenitu sekundarnu strukturu proteinskih slijedova s više od dvije vrste sekundarne strukture.

3.1.5. Fizikalna interpretacija nasumičnih modela strukture proteina - veza s entropijom

Koristeći prethodno izvedeni izraz za mogući broj konformacija (realizacija) modelne strukture proteinskog lanca za slučaj u kojemu ne razlikujemo segmente, mogla bi se izračunati i modelna konformacijska entropija za modele strukture proteinskih lanaca kojima je riješena struktura. Promatramo li entropiju kao mjeru neuređenosti (ali i kompleksnosti tj. složenosti) sustava, mogla bi se definirati modelna konformacijska entropija za lance prema izrazu za slučaj ne-razlikovanja segmenata i reći da je kompleksnost strukture veća što je modelna konformacijska entropija veća. Drugačije rečeno, lanci koji imaju veću modelnu konformacijsku entropiju teže se predviđaju algoritmima.

S ciljem dobivanja što je moguće boljeg modela za predviđanje strukture proteinskih lanaca (tj. položaja segmenata) nastoji se dobiti reprezentativni skup koji će imati najveću modelnu konformacijsku entropiju, jer je takav skup najteži (najsloženiji, najzahtjevniji) za predvidjeti. Tada se u algoritmima za izbor reprezentativnog skupa nastoje pronaći takvi proteinski lanci čije će strukture imati najveću zbirnu (ukupnu) modelnu konformacijsku entropiju:

$$S = \ln(W)$$

Dodatno, u eksperimentalnim podacima za položaje TM segmenata često imamo razmak između pojedinih TM segmenata. Stoga se dodatno definira vrijednost modelne konformacijske entropije

S_i $(i = 1, 2, 3, \ldots)$ koja je jednaka logaritmu (po prirodnoj bazi) broja svih mogućih modelnih konformacija strukture za pojedini lanac, pri čemu je i najmanji mogući razmak između pojedinih segmenata.

Ovdje je potrebno napomenuti da ovako definirana modelna konformacijska entropija strukture nije isto što i termodinamička entropija proteinskog lanca niti konformacijska entropija usljed preraspodjele konformacije aminokiselina u proteinskom lancu.Naime, ovako definirana entropija je entropija preraspodjele TM segmenata na modelnoj strukturi proteinskog lanca s pridruženim sekundarnim strukturama.

Može se analizirati za koji će postotak broja aminokiselina u membrani (u odnosu na ukupni broj aminokiselina unutar proteinskog slijeda) ova konformacijska entropija modelne strukture imati maksimum. Jednostavni problem strukture lanca s N aminokiselina od kojih n može biti u stanju 1, a njih $(N - n)$ u stanju 2 analogan je problemu razmještanja n čestica koje raspoređujemo u N kutija. U tom slučaju za očekivati je (intuitivno) da će se maksimalni broj modelnih konformacija (tj. realizacija) strukture, a time i maksimalna modelna konformacijska entropija strukture (entropija najvjerojatnijeg stanja), dobiti kada je $n = N/2$, što zapravo odgovara binomnoj raspodjeli, a izvod (dokaz) navodimo u nastavku.

3.1.5.1. Binomni nasumični model (izvod izraza za najvjerojatnije stanje)

Pretpostavimo da imamo N aminokiselina u proteinskom slijedu koji mogu zauzeti jedno od dva stanja ('u membrani' ili 'izvan membrane'). Pitamo se za koju će vrijednost aminokiselina u membrani (n_1) i izvan nje (n_2) broj mogućih realizacija u binomnom nasumičnom modelu imati maksimum. Koristeći izraz za broj mogućih ralizacija binomnog nasumičnog modela

$$W = \frac{N!}{n_1! \cdot n_2!}$$

i vjerojatnosti pojedinih stanja

$$p_1 = \frac{n_1}{N} \text{ i } p_2 = \frac{n_2}{N},$$

te logaritma po prirodnoj bazi $\ln(W)$ broja mogućih realizacija dobiva se

$$\ln(W) = \ln\left(\frac{N!}{n_1! \cdot n_2!}\right) = n_1 \ln\left(\frac{N}{n_1}\right) + n_2 \ln\left(\frac{N}{n_2}\right) = -N(p_1 \ln p_1 + p_2 \ln p_2),$$

Uz zamjenu $p_2 = (1 - p_1)$ dobiva se izraz:

$$\ln(W) = -N \cdot [p_1 \ln p_1 + (1 - p_1) \ln (1 - p_1)],$$

$$\frac{\partial}{\partial p_1}\left[\frac{\ln(W)}{N}\right] = \frac{\partial}{\partial p_1}\{-[p_1 \ln p_1 + (1 - p_1) \ln (1 - p_1)]\} = \ln\left(\frac{1 - p_1}{p_1}\right) = 0,$$

tj. ako je:

$$\frac{1 - p_1}{p_1} = 1 \Rightarrow p_1 = \frac{1}{2} \Rightarrow n_1 = n_2 = \frac{N}{2}.$$

Za slučaj kada se aminokiselinama u lancu s ukupno N aminokiselina nasumično pridružuje n stanja sekundarne strukture α (a aminokiselinama na $N - n$ mjesta nepravilna sekundarna struktura), maksimalni broj realizacija strukture s dva stanja prema binomnoj raspodjeli dobiva se za $n = \mathrm{N}/2$:

$$S\left(\frac{N}{2}\right) = ln\left[W\left(\frac{N}{2}\right)\right] = ln\left(\frac{N!}{\frac{N}{2}!\cdot\frac{N}{2}!}\right) = ln(N!) - 2ln\left(\frac{N}{2}!\right).$$

Koristeći Stirlingovu aproksimativnu formulu

$$n! \sim \sqrt{2\pi n}\cdot\left(\frac{n}{e}\right)^n,$$

dobiva se

$$S\left(\frac{N}{2}\right) = ln\left[\sqrt{2\pi N}\left(\frac{N}{e}\right)^N\right] - 2ln\left[\sqrt{2\pi\frac{N}{2}}\left(\frac{\frac{N}{2}}{e}\right)^{\frac{N}{2}}\right] = Nln2 - \frac{1}{2}ln\left(\frac{\pi N}{2}\right).$$

Iz ovoga slijedi da, kada je N velik (tada je $\ln N \ll N$), entropija (tj. maksimalni broj realizacija) koja se dobije za $n = N/2$ a koja se izračunava prema aproksimativnom izrazu uporabom Strilingove formule, zanemarivo malo razlikuje od točnog izračuna entropije koja je dana izrazom:

$$S = \ln(W) = \ln(2^N) = N\ln 2\,.$$

U tablici 15. prikazane su vrijednosti raspodjele broja realizacija u primjeru binomnog nasumičnog modela strukture proteina za slijedove/lance duljine $N = 5$ i $N = 8$.

Tablica 15. Broj kombinacija i vjerojatnosti stanja binomnih koeficijenata za duljine slijeda $N = 5$ i $N = 8$.

$N = 5$			
# zauzetih kutija u slijedu	% zauzetih kutija u slijedu	broj kombinacija	vjerojatnost
0	0	1	0.0312
1	20	5	0.1562
2	40	10	0.3125
3	60	10	0.3125
4	80	5	0.1562
5	100	1	0.0312
	ukupno	$32 = 2^5$	1

$N = 8$			
# zauzetih kutija u slijedu	% zauzetih kutija u slijedu	broj kombinacija	vjerojatnost
0	0	1	0.0039
1	12.5	8	0.0313
2	25	28	0.1094
3	37.5	56	0.2188
4	50	70	0.2734
5	62.5	56	0.2188
6	75	28	0.1094
7	87.5	8	0.0313
8	100	1	0.0039
	ukupno	$256 = 2^8$	1

Ukupni broj mogućih kombinacija jednak je 2^N, dok je vjerojatnost raspodjele određenog broja aminokiselina u stanju 1 (n) u slijedu duljine N jednak broju kombinacija za to stanje podijeljeno s ukupnim brojem kombinacija za sva moguća stanja. Ako je N neparan, najvjerojatnija su ona stanja (stanja s najvećim brojem kombinacija/realizacija) koja su najbliže $N/2$ (u slučaju $N = 5$ to su stanja za $n = 2$ i $n = 3$). Za parne vrijednosti N, najvjerojatnije stanje-točno je za $\frac{N}{2}$ (u slučaju $N = 8$ to je za $n = 4$). Ako je N paran broj uočavamo da se tada maksimum nalazi točno na $n = \frac{N}{2}$, a ako je N neparan imamo dva jednaka rješenja koja se nalaze na položajima $n_1 = (\mathrm{N} - 1)/2$ i $n_2 = (\mathrm{N} + 1)/2$. Broj kombinacija pri polovičnom zauzeću $(n = \mathrm{N}/2)$ svakako raste s brojem N, jer vrijede nejednakosti:

$$\binom{N+1}{\frac{N}{2}} > \binom{N}{\frac{N}{2}} - \; za\ parno\ N \in \mathbb{N}$$

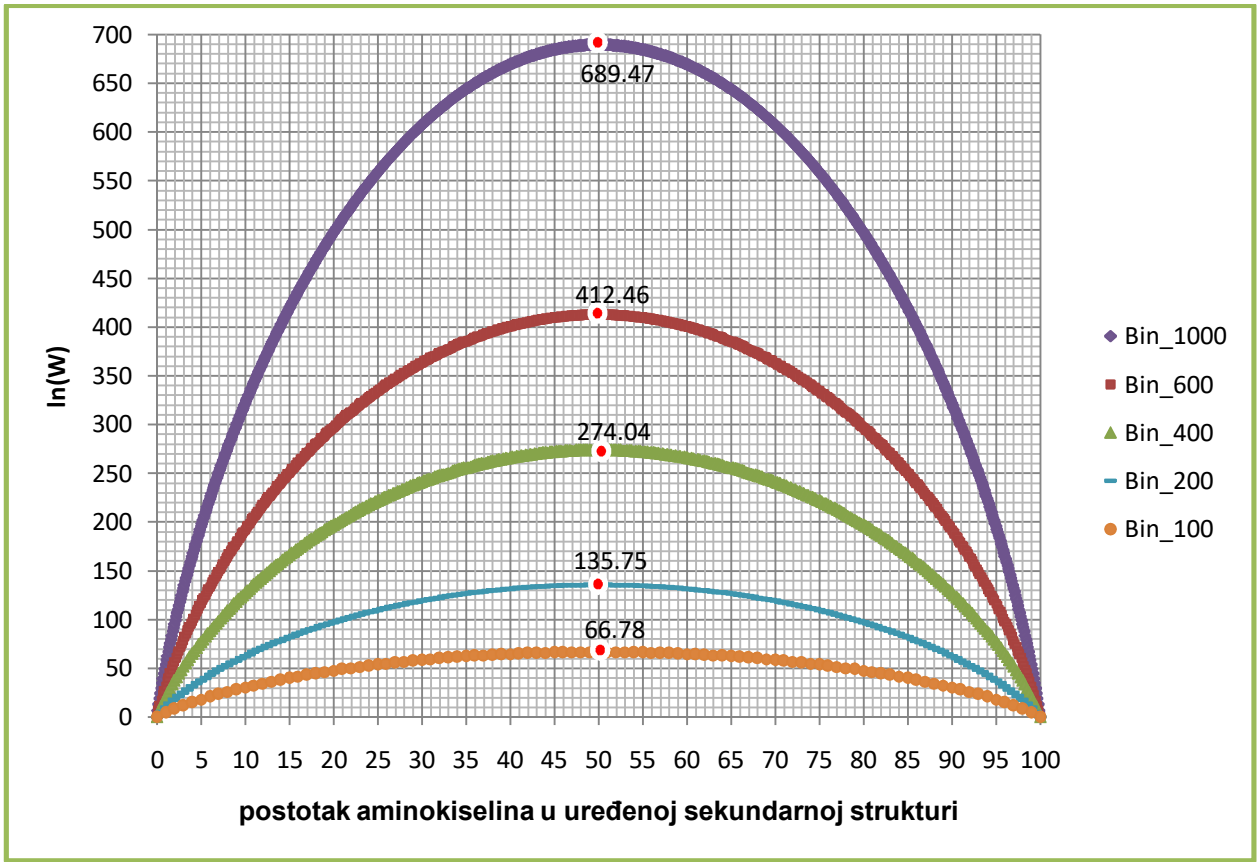
$$\binom{N+1}{\frac{N+1}{2}} > \binom{N}{\frac{N-1}{2}} - \; za\ neparno\ N \in \mathbb{N}$$

Broj kombinacija (realizacija strukture lanca, tj. broj mogućih rasporeda) za pojedina stanja u ovisnosti o postotnom udjelu broja aminokiselina u stanju 1 (n), kao i pripadajuće vjerojatnosti, simetrični su u odnosu na $(n = N/2)$. Taj je rezultat i očekivan, jer za binomne koeficijente vrijedi

$$\binom{N}{k} = \binom{N}{N-k}$$

Za svako N najveći broj kombinacija dobiva se kada je odnos zauzetih i nezauzetih stanja 50:50, i on ne ovisi o promjeni duljine slijeda N. Za slučaj kada je N neparan, imamo dva maksimuma za vrijednosti koje su najbliže (s jedne i s druge strane) stanju 50:50.

Na slici 16. prikazan je logaritam broja mogućih realizacija (modelnih konformacija) binomnog nasumičnog modela u ovisnosti o postotku aminokiselina u sekundarnoj strukturi za proteinske slijedove duljina 100, 200, 400, 600 i 1000. Vidi se kako su za sve slijedove maksimumi (označeni crveno ispunjenim krugom) kada je polovica $(n = 50\%)$ aminokiselina u uređenoj sekundarnoj strukturi. Dodatno, vidimo da maksimalni broj nasumičnih realizacija modelne sekundarne strukture raste s porastom duljina proteinskog slijeda.



Slika 16. Ovisnost logaritma broja mogućih realizacija (modelnih konformacija) binomnog nasumičnog modela o postotku aminokiselina u sekundarnoj strukturi (prikaz za proteinske slijedove duljina 100, 200, 400, 600 i 1000).

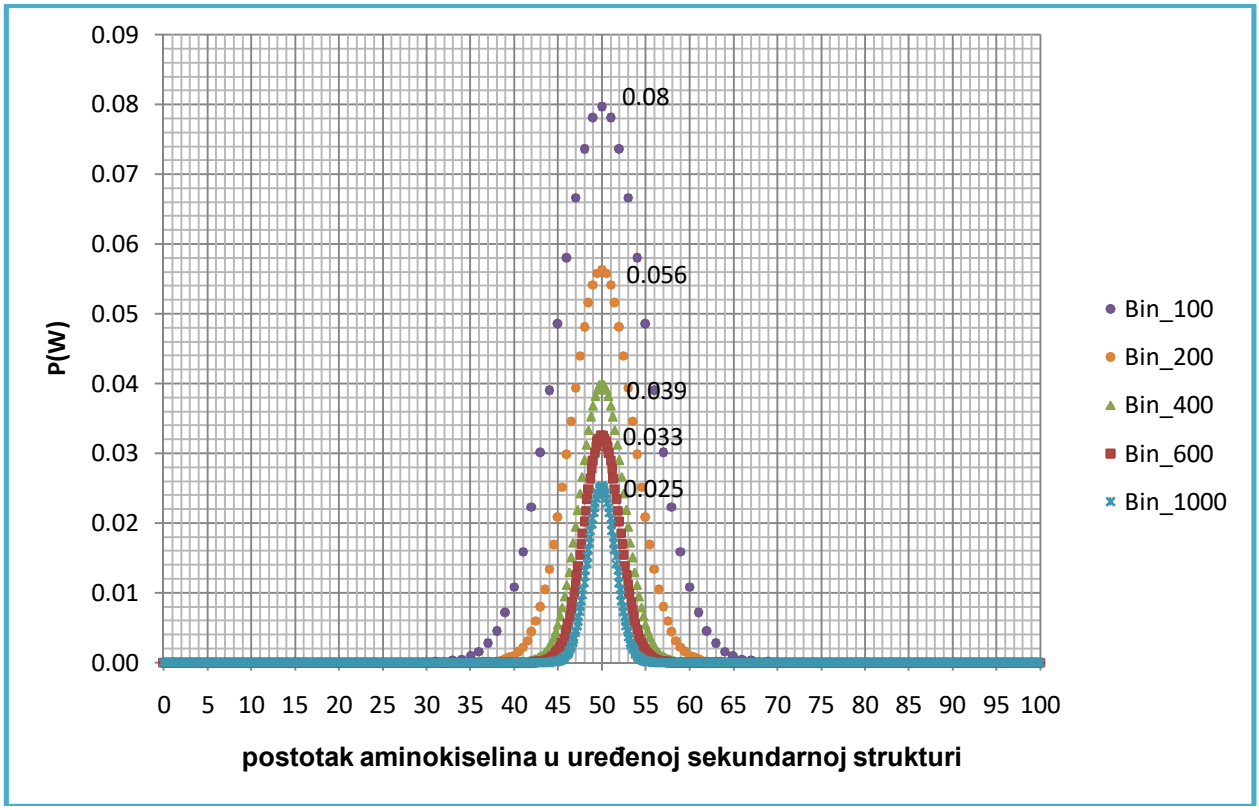
Međutim, ako se za pojedini lanac s N aminokiselina promatraju raspodjele vjerojatnosti stanja u modelu s dvije sekundarne strukture (pritom je zbroj vjerojatnosti svih stanja jednak 1) tada vjerojatnost stanja s maksimalnim brojem konformacija opada s porastom N, što je prikazano na slici 16. Važno je imati na umu da su sve vrijednosti raspodjele (za sve lance) na slici 16. diskretne a ne kontinuirane. Tako je gustoća stanja raspodjele sve veća kada idemo od kraćih ka dužim lancima.

Omjeri maksimalnih vrijednosti logaritama broja mogućih realizacija binomnog nasumičng modela za duljine N_1 i N_2 definirani su izrazom:

$$\frac{S\left(\frac{N_1}{2}\right)}{S\left(\frac{N_2}{2}\right)} = \frac{N_1 ln2 - \frac{1}{2} ln\left(\frac{\pi N_1}{2}\right)}{N_2 ln2 - \frac{1}{2} ln\left(\frac{\pi N_2}{2}\right)}$$

što za duljine $N_1 = k \cdot N$ i $N_2 = N$ za velike N daje:

$$\lim_{N\to\infty}\left[\frac{S\left(\frac{kN}{2}\right)}{S\left(\frac{N}{2}\right)}\right] = \lim_{N\to\infty}\left[\frac{kNln2 - \frac{1}{2} ln\left(\frac{\pi kN}{2}\right)}{Nln2 - \frac{1}{2} ln\left(\frac{\pi \cdot kN}{2}\right)}\right] = k$$



Slika 17. Vjerojatnost realizacija (modelnih konformacija) binomnog nasumičnog modela u ovisnosti o postotku aminokiselina u sekundarnoj strukturi (prikaz za proteinske slijedove duljina 100, 200, 400, 600 i 1000).

Stanje s maksimalnim brojem realizacija za $N = 600$ na slici 17 ima manju vjerojatnost (0.033) u raspodjeli svih stanja toga lanca nego li stanje s maksimalnim brojem realizacija za lanac $N = 400$ (0.039). Razlog tome je što ukupni zbroj vjerojatnosti mora biti jednak jedinici, a raspodjela Bin_400 (za lanac $N = 400$) ima 400 diskretnih vrijednosti, dok raspodjela Bin_600 (za lanac $N = 600$) ima 600 takvih (vrijednosti) vjerojatnosti. Stoga, doprinos stanja maksimalne vjerojatnosti u raspodjeli za lanac $N = 600$ manji je (ili, ima manju težinu ~ 1/600) nego doprinos stanja maksimalne vjerojatnosti za lanac $N = 400$ (težina ~ 1/400). Analogno vrijedi i za druge raspodjele sa slike 17. Gustoća stanja igra bitnu ulogu kada se razmatraju ili uspoređuju raspodjele koje odgovaraju različitim dužinama lanaca.

Općenito govoreći, vjerojatnost najvjerojatnije raspodjele, tj. samog vrha šiljka jednaka je:

$$P\left(\frac{N}{2}\right) = \frac{1}{2^N} \cdot \binom{N}{\frac{N}{2}} = \frac{1}{2^N} \cdot \frac{N!}{\left(\frac{N}{2}\right)! \cdot \left(\frac{N}{2}\right)!}$$

što uz Strilingovu formulu daje

$$P\left(\frac{N}{2}\right) = \frac{1}{2^N} \cdot \frac{ln\left[\sqrt{2\pi N}\left(\frac{N}{e}\right)^N\right]}{\left[\sqrt{2\pi\frac{N}{2}}\left(\frac{\frac{N}{2}}{e}\right)^{\frac{N}{2}}\right] \cdot \left[\sqrt{2\pi\frac{N}{2}}\left(\frac{\frac{N}{2}}{e}\right)^{\frac{N}{2}}\right]} = \sqrt{\frac{2}{\pi N}}$$

Dakle, za vrlo veliki *N*, vjerojatnost određene razdiobe (stanja), pa tako i najvjerojatnije, postaje sve manja. Pitanje je dakle, koliko ima razdioba u vrhu šiljka, koje iscrpljuju 'gotovo sve', moguće realizacije. Grubi odgovor je, upravo $\sqrt{N\pi/2}$. Naime, ako sve razdiobe u 'tjemenu' (pri čemu se misli na $\sqrt{N\pi/2}$ razdioba koje su najbliže 'tjemenu') imaju podjednaku vjerojatnost, a zbroj njihovih vjerojatnosti približno je 1, onda njih ima baš onoliko koliko iznosi recipročna vrijednost te vjerojatnosti. Ovaj odgovor je približno točan. Obzirom da su vrijednosti razdioba bliske 'šiljku' čija je vjerojatnost približno jednaka $\sqrt{2/(N\pi)}$, to treba uzeti približno $\sqrt{N\pi/2}$ razdioba. Kako ima ukupno *N* razdioba to je udio razdioba proporcionalan $\sqrt{N}/N$ što za veliki *N* teži nuli (odnosno zanemarivo mali broj u odnosu na ukupni broj razdioba [72]

Nadalje, postavlja se pitanje koliki će interval (izražen u postotku aminokiselina u stanju 1, tj. u sekundarnoj strukturi α) u binomnoj raspodjeli obuhvatiti npr. 99% (značajna vjerojatnost intervala $\pm 2.58\sigma$) svih stanja (slika 17). Za lanac duljine $N = 100$ taj je interval (granice su uključene) između 38 i 62, za $N = 200$ interval je između 40 i 60, a za $N = 1000$ između 46 i 54. Vidi se kako porastom *N* taj interval postaje sve uži oko srednje (maksimalne) vrijednosti (vjerojatnosti), a razlog za sužavanje tog intervala je sve gušća raspodjela. Za jako velike *N* smatra se da je taj interval toliko uzak da su skoro sva stanja značajnije vjerojatnosti oko 50:50 posto (tj. oko $n = N/2$), dok su druga stanja udaljenija od sredine raspodjele (maksimuma) zanemarive vjerojatnosti, što objašnjava standardni model raspodjele čestica u prostoru. Uz pretpostavku da imamo dvije jednake spojene posude ispunjene plinom mi mjerimo jednake tlakove kao raspodjelu jednakog broja čestica u posudama (pri istoj temperaturi). Ako u posudi imamo na primjer $N = 10^{20}$ molekula plina, čak i da imamo u jednoj od posuda npr. 10000 molekula više, makroskopski to nećemo mjeriti jer je to zanemarivo mala razlika. Dakle iako smo se odmakli od točnog odnosa broja čestica u zdjelama 50:50, vjerojatnost tog stanja i dalje je jako visoka i plin će se nalaziti u nekom od ovih približnih stanja.

3.1.5.2. Segmentni nasumični model (izvod zraza za najvjerojatnije stanje)

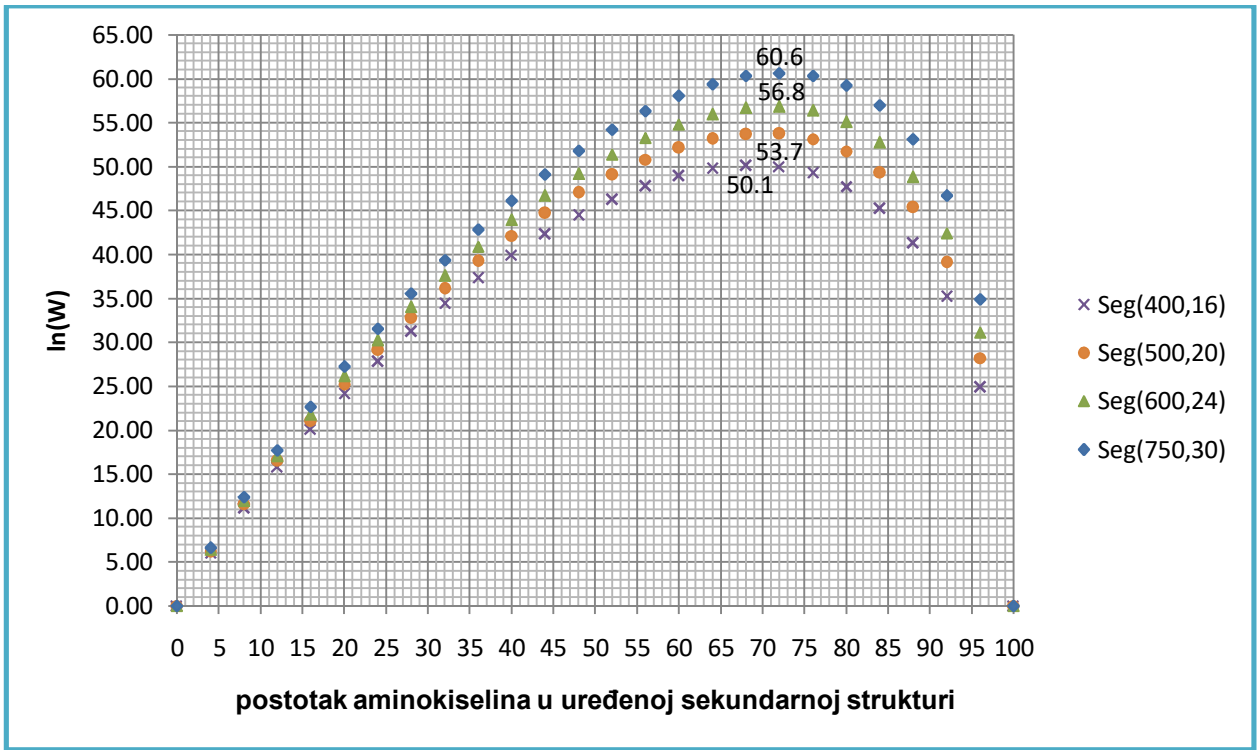
Postavlja se pitanje - hoće li u slučaju aminokiselinskog slijeda duljine *N* s dvije vrste sekundarne strukture (dva stanja) maksimum modelne konformacijske entropije za segmentni nasumični model biti za slučaj kada je podjednak broj aminokiselina ($n = N/2$) u svakom od stanja? U tom prikazu strukturu lanca čine segmenti istoga stanja, različite nasumične realizacije strukture mogu se dobiti samo različitim raspoređivanjem tih segmenata na preostale dopuštene položaje unutar lanca (a ti se položaji nalaze u drugom stanju, tj. sekundarnoj strukturi i ima ih $(N - n)$).

Na slici 18. prikazana je usporedna analiza logaritma broja mogućih realizacija modelne strukture prema segmentnom nasumičnom modelu za slijedove duljine (400, 500, 600, 750) u ovisnosti o postotku aminokiselina u lancu koje poprimaju pravilnu sekundarnu strukturu. Raspodjele sačinjavaju duljine slijedova koje pri maksimalnoj popunjenosti imaju točno određeni broj segmenata, u ovom slučaju 25 kod svake raspodjele. Slijed duljine 400 ima sve segmente duljine 16 (tako da je maksimalno $400/16 = 25$ segmenata), a označavamo ga kao Seg(400,16), i analogno se označavaju i drugi razmatrani slijedovi: (500,20), (600,24) i (750,30). Svaka raspodjela prikazuje broj stanja tako da je broj točaka u svakoj raspodjeli podjednak (25 diskretnih vrijednosti), kao i njihova gustoća. Primjerice, za slijed duljine 400 računa se broj mogućih realizacija modelne strukture za sve segmente duljine 16 polazeći od jednog segmenta u lancu do stanja kada je u lancu 25 segmenata. Analogno je provedena analiza za slijedove duljine 500, 600 i 750 aminokiselina. Primjeri lanaca izabrani su tako da:

(a) duljine segmenata budu blizu prosjeka stvarne duljine TM segmenta u integralnim membranskim slijedovima,

(b) da se u svaki od četiri odabrana slijeda (400, 500, 600 i 750) može smjestiti 1 do 25 segmenata, što je raspon broja TM segmanata u integralnim membranskim proteinima (alfa vrste) kojima je određena struktura.

Pokazuje se da su postotci aminokiselina u uređenoj sekundarnoj strukturi pri kojima se dobivaju maksimalni brojevi realizacija modelnih struktura u segmentnom nasumičnom modelu (za sve duljine slijedova) uvijek iznad 65%. To je različito od binomnog (ne-segmentnog) nasumičnog modela strukture kod kojega su maksimumi uvijek točno na 50%, tj. kada je u lancu polovica aminokiselina ($n = N/2$) u jednom, a druga polovica u drugom stanju. Maksimalna je vrijednost broja realizacija to veća što je slijed duži, i pri tome maksimum se pojavljuje pri većem postotku aminokiselina u stanju 1 (u TM segmentima, odnosno u sekundarnoj strukturi α).



Slika 18. Ovisnost logaritma broja mogućih realizacija (modelnih konformacija) segmentnog nasumičnog modela o postotku aminokiselina u uređenoj sekundarnoj strukturi (prikaz za proteinske slijedove duljina 400, 500, 600 i 750).

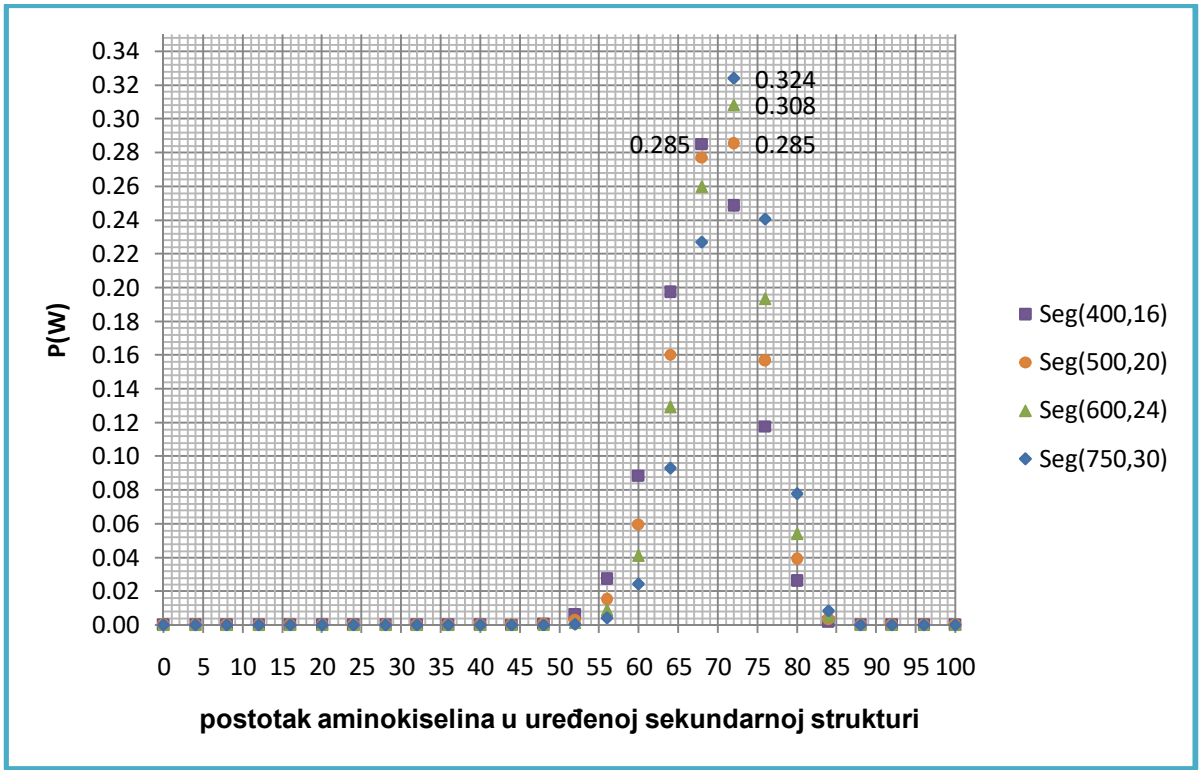
Maksimalni broj realizacija modelne strukture u segmentnom nasumičnom modelu za duljinu slijeda $N = 400$ - $Seg(400,16)$ dobiva se pri popunjenosti proteinskog slijeda (tj. postotku aminokiselinama u jednom stanju strukture – npr. α vrste) od 68%. Za slijedove duljina 500, 600 i 750, maksimumi se pojavljuju na popunjenosti od 72%. Moguće je dobiti i dva maksimuma broja realizacija za dva postotka popunjenosti, a razlika među njima može se izračunati tako da 100 podijelimo s maskimalnim brojem mogućih segmenata (pri maksimalnoj popunjenosti, tj. kad su sve aminokiseline u segmentima). U ovom primjeru to je $100/25 = 4$, i maskimumi broja realizacija su pri popunjenostima jednom vrstom sekundarne strukture od 68% i 72%. Kako svaki segment reducira preostali dio slijeda za $(d - 1)$, bilo je i za očekivati da se u ovakvom modelu maksimumi raspodjele neće nalaziti na polovičnoj popunjenosti.

Pogledajmo sada kolike su vjerojatnosti stanja u kojemu imamo maksimum raspodjele, tj. maksimum broja realizacije modelne strukture u segmentnom nasumičnom modelu, i kako te vjerojatnosti ovise o promjeni duljine slijeda (uz zadržavanje jednakog odnosa između duljine slijeda i duljine duljina segmenta).

Na slici 19 uočava se da vjerojatnosti stanja u kojima imamo najveći broj realizacija raste s porastom duljine slijeda. Za proteinski lanac s $N = 750$ i segmente od 30 aminokiselina Seg(750,30), vjerojatnost stanja s maksimalnim brojem realizacija iznosi 0.32.

Nadalje, pogledajmo u kojem je intervalu oko maksimuma segmentne nasumične raspodjele zbroj vjerojatnosti članova veći od 0.99 (analogija sa normalnom raspodjelom za interval $\pm 2.58\sigma$ oko srednje vrijednosti). Za slučaj Seg(400,16), $s_{max} = 17$ taj je interval $[s_{max} - 4, s_{max} + 4] \equiv [13,21]$, dok je za slučaj Seg(750,30), $s_{max} = 18$ taj interval u rasponu $[s_{max} - 3, s_{max} + 3] \equiv [15,21]$. Dakle, interval oko maksimuma segmentne nasumične raspodjele u kojemu je zbroj vjerojatnosti članova veći od 99% sužava se povećanjem duljine proteinskog slijeda (uz konstantni omjer duljine slijeda i duljine TM segmenta).

To se slaže s onim što se dobiva kod binomnog nasumičnog modela gdje se interval, u kojem se interval dominantnih razdioba nasumičnih modelnih realizacija sužava s porastom N (za izrazito veliki N praktički sve značajne razdiobe binomne modelne strukture budu na podjednakoj zastupljenosti (popunjenosti) jedne i druge vrste sekundarne strukture u lancu 50:50). Kako se interval s dominantnim brojem razdioba u segmentnom nasumičnom modelu sužava s porastom N, i sam broj značajnih razdioba modelnih struktura za određenu popunjenost slijeda (ili broj segmenata u lancu) opada proporcionalno sužavanju intervala. Naime, ukupni broj razdioba (gustoća stanja) u segmentnom nasumičnom modelu je stalan i jednak broju TM segmenata, što je suprotno od binomnog nasumičnog modela gdje gustoća stanja raste proporcionalno duljini slijeda. Kako porastom duljine slijeda N dominantni dio raspodjele u segmentnom nasumičnom modelu postaje sve uži, a gustoća stanja je konstantna to maksimumi raspodjele postaju u postotku sve viši (jer ukupni zbroj vjerojatnosti značajnih razdioba treba biti ~ 1) što je također u suprotnosti sa vjerojatnostima za binomnu nasumičnu raspodjelu gdje vjerojatnost maksimuma opada s porastom duljine slijeda N.



Slika 19. Vjerojatnost realizacija (modelnih konformacija) segmentnog nasumičnog modela u ovisnosti o postotku aminokiselina u sekundarnoj strukturi (prikaz za proteinske slijedove duljina 400, 500, 600 i 750).

Postavlja se pitanje, što se događa ako se u slijedu duljine *N*, dodaje proizvoljni broj segmenata točno određenih (i jednakih) duljina? U tom se slučaju broj stanja računa prema izrazu:

$$W(N,d,s) = \binom{N+s-ds}{s} = \binom{N-(d-1)s}{s},$$

gdje je *N* – duljina slijeda, *d* – duljina segmenta, a *s* – broj segmenata u modelnoj strukturi proteinskog lanca. Kako bi se pronašao član sa maksimalnim brojem realizacija modelne strukture u izrazu za $W(N,d,s)$, zbog faktorijela u gornjem izrazu, primjenjuje se Stirlingova formula na logaritam broja stanja (konformacijska entropija), što vodi na:

$$\ln[W(N,d,s)] = \ln\binom{N+s-ds}{s} = \ln\binom{N-(d-1)s}{s} = \ln\frac{[N-(d-1)s]!}{s!\,(N-ds)!}$$
$$= [N-(d-1)s]\ln[N-(d-1)s] - s\ln s - (N-ds)\ln(N-ds).$$

Ako uzmemo određenu (fiksnu) duljinu slijeda *N* i duljinu segmenta *d*, ova funkcija ovisit će samo o broju segmenata *s*. Maksimum ove funkcije u ovisnosti o broju segmenata nalazi se u točki gdje je prva derivacija *s* jednaka nuli (ovdje se ostavlja oznaka za parcijalnu derivaciju kako bi se naglasilo da je opća ovisnost o *N, d, s,* a ne samo o *s*). To nas vodi na jednadžbu:

$$\frac{\partial \ln[W(N,d,s)]}{\partial} = \frac{\partial}{\partial s}\{[N-(d-1)s]\ln[N-(d-1)s] - s\ln s - (N-ds)\ln(N-ds)\}$$
$$= -(d-1)\ln[N-(d-1)s] - \ln s + d\ln(N-ds) = 0$$

odnosno na

$$d\ln\left(1+\frac{s}{N-ds}\right) = \ln\left(1+\frac{N-ds}{s}\right),$$

koja se još može napisati i u obliku

$$d\ln\left(\frac{\frac{N}{s}-d+1}{\frac{N}{s}-d}\right) = \ln\left(\frac{N}{s}-d+1\right).$$

Uvodeći zamjenu:

$$\frac{N}{s}-d+1 = x$$

dobiva se:

$$d\ln\left(\frac{x}{x-1}\right) = \ln x$$

Ova vrsta jednadžbe nije egzaktno rješiva. Stoga se pristupilo traženju aproksimativnoga rješenja koje će dati zadovoljavajuću procjenu položaja maksimuma raspodjele broja realizacija modelne strukture u segmentnom nasumičnom modelu. Aproksimativna funkcija dobivena je tražeći takav omjer (N/s_{max}) za koji će se dobiti najveći broj mogućih realizacija (ovdje je *N* duljina slijeda, a s_{max} broj segmenata u modelnoj strukturi koja daje maksimalni broj realizacija).

Analizom slijedova duljina 1000 i 500, došlo se do rezultata prikazanih u tablici 16.

Tablica 16. Analiza broja segmenata koji daju najveći broj realizacija za duljine slijedova (1000 i 500) i odnosa N/s_{max}.

$N = 1000$			$N = 500$		
d	s_{max}	N/s_{max}	d	s_{max}	N/s_{max}
1	500	2.0	1	250	2.0
2	276	3.6	2	138	3.6
3	194	5.2	3	97	5.2
4	151	6.6	4	75	6.7
5	124	8.1	5	62	8.1
6	105	9.5	6	53	9.4
7	92	10.9	7	46	10.9
8	81	12.3	8	41	12.2
9	73	13.7	9	36	13.9
10	66	15.2	10	33	15.2
16	43	23.3	16	22	22.7
20	35	28.6	20	18	27.8
50	15	66.7	50	7	71.4
100	8	125.0	100	4	125
200	4	250.0	200	2	250
$pkk(d, N/s_{max}) = 0.9999$			$pkk(d, N/s_{max}) = 0.9995$		

N – duljina slijeda
d – duljina segmenta
s_{max} – broj segmenata u sekundarnoj strukturi slijeda za koju se dobiva maksimalni broj nasumičnih realizacija modelne strukture
pkk – Pearsonov koeficijent korelacije

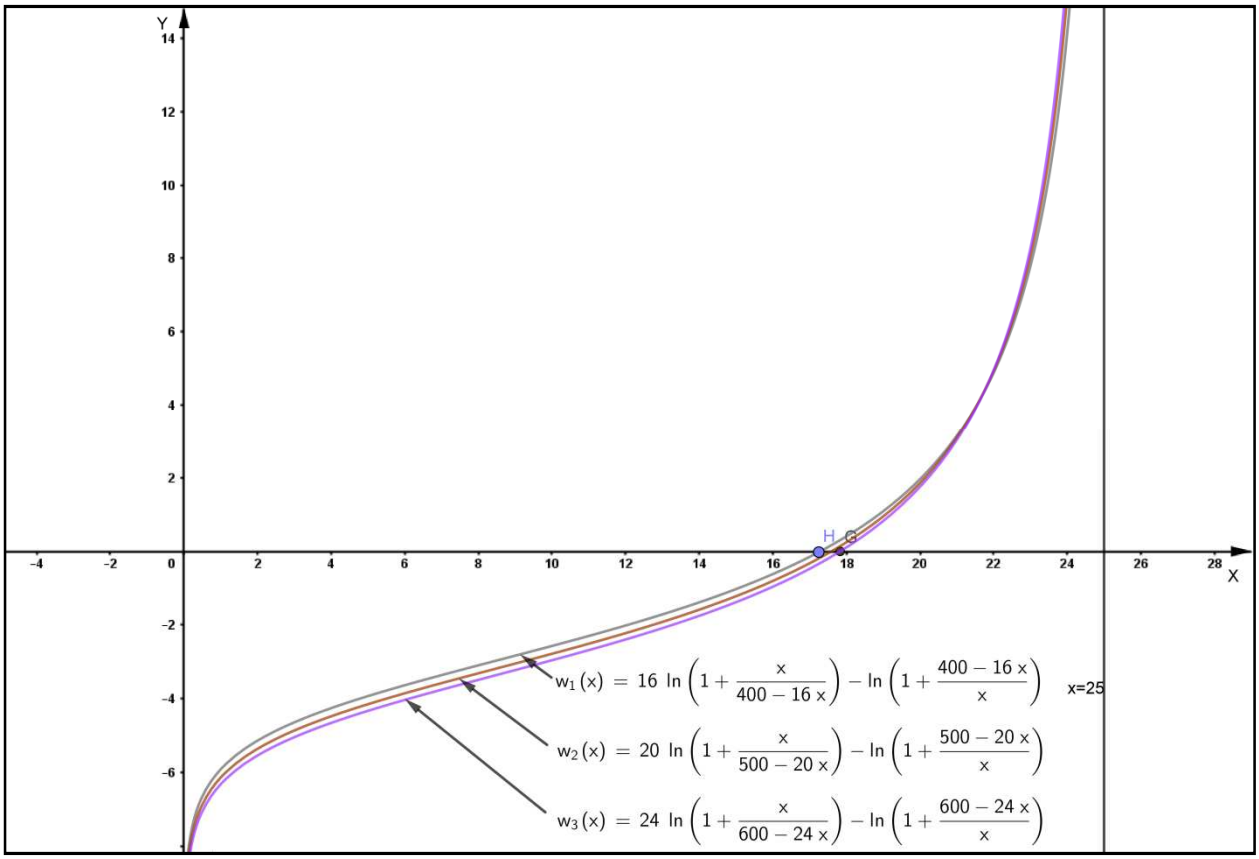
Kao što se može vidjeti odnos N/s_{max} za slijedove 1000 i 500 pokazuju dosta dobro slaganje. Na osnovu analize dobivenih maksimuma za različite vrijednosti duljina slijeda i segmenata došlo se do izraza za broj segmenata koji će davati (uz konstantni omjer duljine slijeda i duljine segmenata najveći broj mogućih realizacija:

$$n_{max} \cong \frac{N}{2} \cdot d^{-\frac{\sqrt{\pi}}{2}}$$

Ovaj se izraz podudara s rezultatom za maksimalni broj realizacija kod binomnog nasumičnog modela (binomni obrazac) koji se dobije za duljina segmenta $d = 1$, uz $n_{max} = N/2$).

Na slici 20. prikazane su analitičke funkcije prve derivacije logaritma broja mogućih nasumičnih realizacija modelne strukture te su tražene njihove nultočke. Kako nultočke ovih funkcija nisu cjelobrojne vrijednosti, gleda se nultočki najbliža cjelobrojna vrijednost. Na taj način dobije se broj segmenata u modelnoj strukturi za koji se dobiva najveći broj nasumičnih realizacija modelne strukture. Rješenje za necjelobrojne vrijednosti značilo bi npr. da smo u slijed ubacili 17 segmenata i jednu četvrtinu 18-og segmenta, što se ne može dogoditi prema definiranom modelu, jer broj segmenata u nasumičnom segmentnom modelu mora biti cjelobrojan. Stoga, u raspodjeli dobivamo maksimum broja mogućih realizacija za modelnu strukturu sa 17 segmenata. Na primjer, za slijed duljine 400 sa segmentima duljina 16 aminokiselina nultočka funkcije prve derivacije logaritma broja nasumičnih realizacija modelne strukture iznosi 17.25 (točka H na slici 20.) Za dvije preostale funkcije nultočka će biti za 18

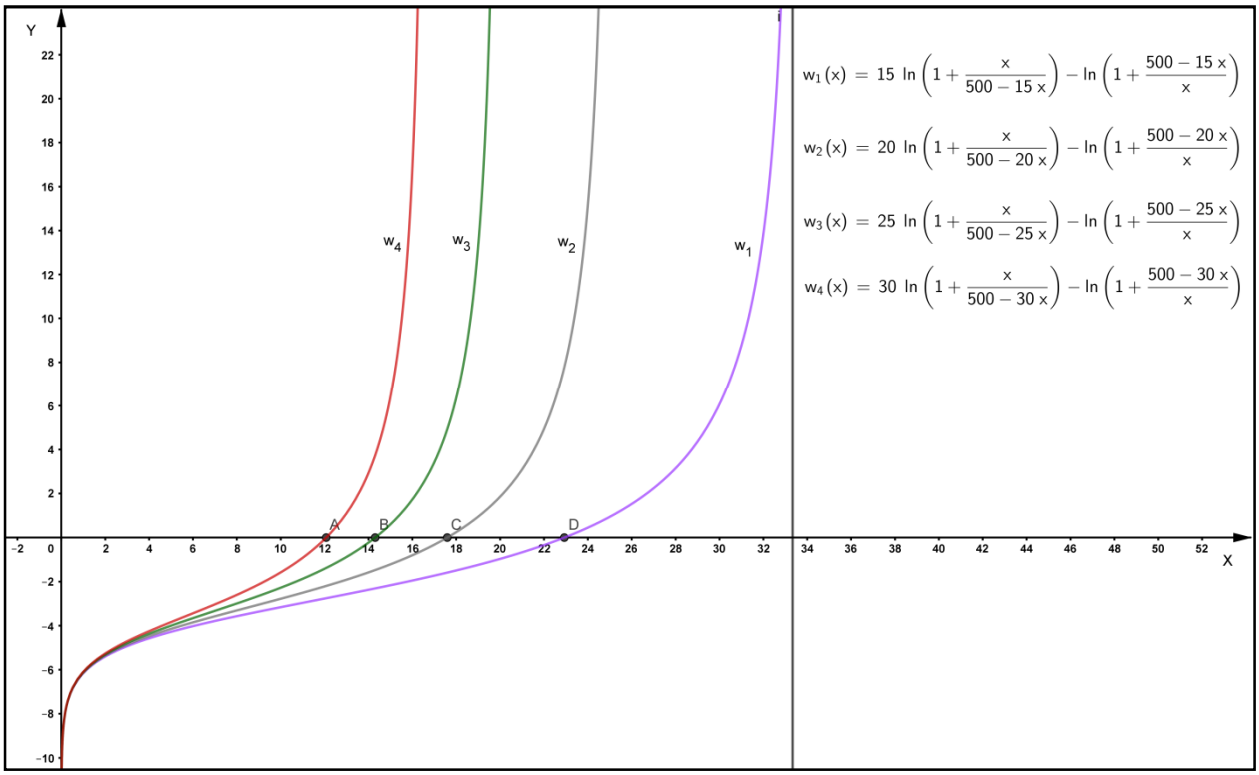
ugrađenih segmenata u slijedu (na slici 20. točka G ima vrijednost 17.83 što zaokruženo na cjelobrojnu vrijednost daje 18 segmenata)



Slika 20. Ovisnost funkcije prve derivacije logaritma broja modelnih konformacija strukture u segmentnom nasumičnom modelu za tri duljine slijedova (400, 500, 600) i (redno) tri duljine segmenata (16, 20, 24) o broju segmenata u lancu.

Asimptotske vrijednosti funkcije prve derivacije logaritma broja nasumičnih realizacija modelne strukture su $x = 0$ i $x = 25$ (što je imaksimalni broj segmenata koji se može ugraditi u slijed). Naime, u izrazu funkcije prve derivacije logaritma broja nasumičnih realizacija modelne strukture za $x = 0$ drugi će član dati ∞ vrijednost, dok za drugu asimptotsku vrijednost $x = 25$ prvi član postaje beskonačan. Porastom duljine segmenata graf funkcije vertikalno se širi, zbog vrijednosti za duljinu segmenta u prvom članu ispred logaritma (ln). Položaji maksimuma broja nasumičnih realizacija modelne strukture (koji se nalaze u nultočki prve derivacije funkcije) neznatno se mijenjaju u odnosu na promjenu duljine slijeda (ili duljine segmenta). Dok se duljina slijeda mijenja od 400 do 600, vrijednosti broja segmenata koji daju maksimalni brojevi nasumičnih realizacija modelne strukture mijenjaju se od 17 do 18 (od točke H do točke G).

U nastavku će se analizirati ovisnost funkcije prve derivacije broja realizacija modelne strukture segmentnog nasumičnog modela kada, umjesto stalnog maksimalno mogućeg broja segmenata koji se mogu smjestiti u slijedu s N aminokiselina, uzmemo da je duljina slijeda nepromijenjena, tj. $N = konst.$. Na slici 21. vidi se da dolazi do širenja grafa u horizontalnom smjeru, pa se raspon vrijednosti u kojima se nalazi maksimum znatnije mijenja u odnosu na relativnu promjenu duljina u prethodnom razmatranju ovisnosti slijed-segment. Uz relativnu promjenu duljine segmenata od 15 do 30 aminokiselina, došlo je do promjene položaja nultočaka prve derivacije (za koje se postiže maksimalni broj nasumičnih realizacija modelne strukture u segmentnom modelu) s 23 (točka D na slici 21) na 12 segmenata (točka A), tj. odnos je negativan.



Slika 21. Ovisnost funkcije prve derivacije logaritma broja modelnih konformacija strukture u segmentnom nasumičnom modelu za duljinu slijeda $N = 500$ i za četiri duljine segmenata (15, 20, 25 i 30) o broju segmenata u lancu.

Položaji vertikalnih asimptota funkcije prve derivacije logaritma broja modelnih konformacija ima onu vrijednost u kojoj nazivnici izraza pod logaritmima imaju nulu. Prvo rješenje (desni član izraza) je nula i kao što se vidi to rješenje je jednako za sve prikazane funkcije. Druga vrijednost ovisi o duljini slijeda i duljini segmenta i jednaka je njihovom količniku, odnosno:

$$x = \frac{N}{d}$$

To npr. za funkciju w_4 na slici iznosi

$$x_4 = \frac{N}{d} = \frac{500}{30} = 16.\dot{6}$$

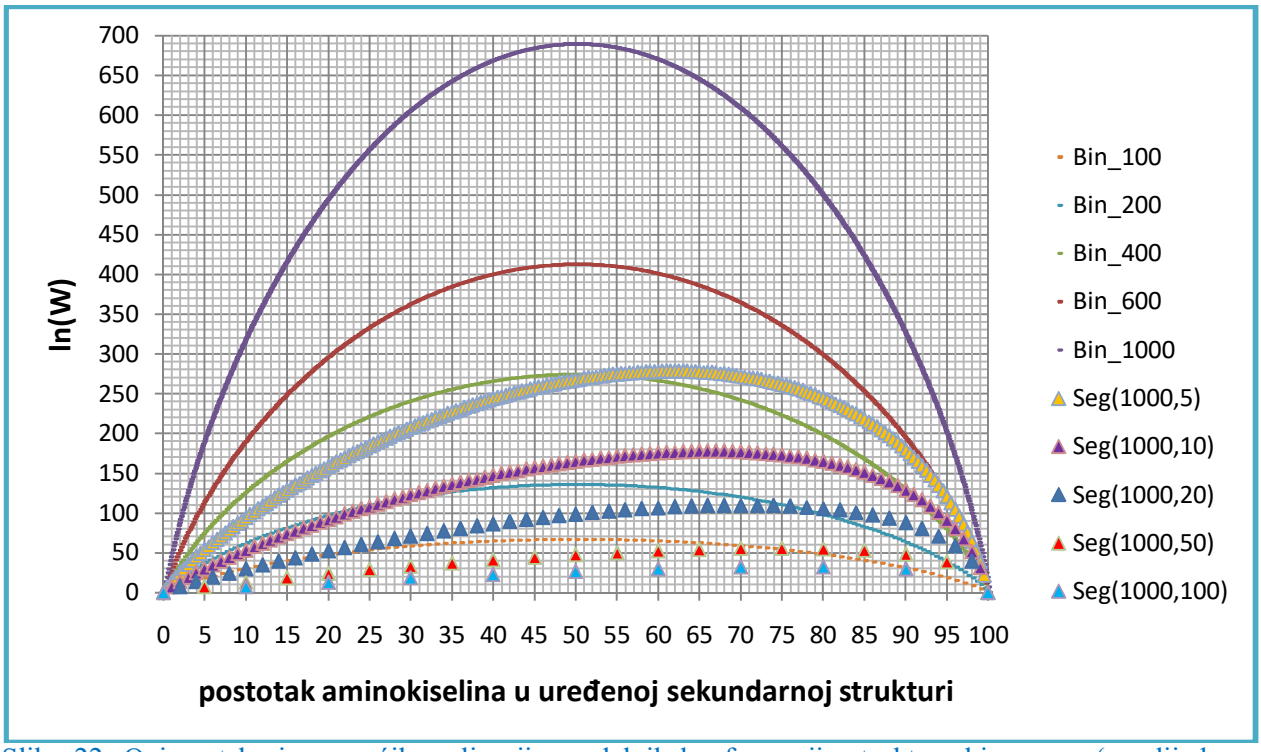
dok je za funkciju w_1

$$x_1 = \frac{N}{d} = \frac{500}{15} = 33.\dot{3}$$

Smanjenjem duljine segmenta d obrnuto proporcionalno se mijenja vrijednost položaja desne vertikalne asimtote, promjenom duljine segmenta sa 30 na 15 (dvostruko smanjenje), desna vertikalna asimptota mijenja vrijednost sa $16.\dot{6}$ na $33.\dot{3}$ (dvostruko veća vrijednost), odnosno graf se širi horizontalno.

3.1.5.3. Usporedba binomnog i segmentnog nasumičnog modela

Usporedba raspodjela vjerojatnosti realizacija kod binomnog i segmentnog nasumičnog modela za potrebu analize najvjerojatnijih stanja (s maksimalnim brojem realizacija modelnih konformacija strukture) dana je na slici 22.



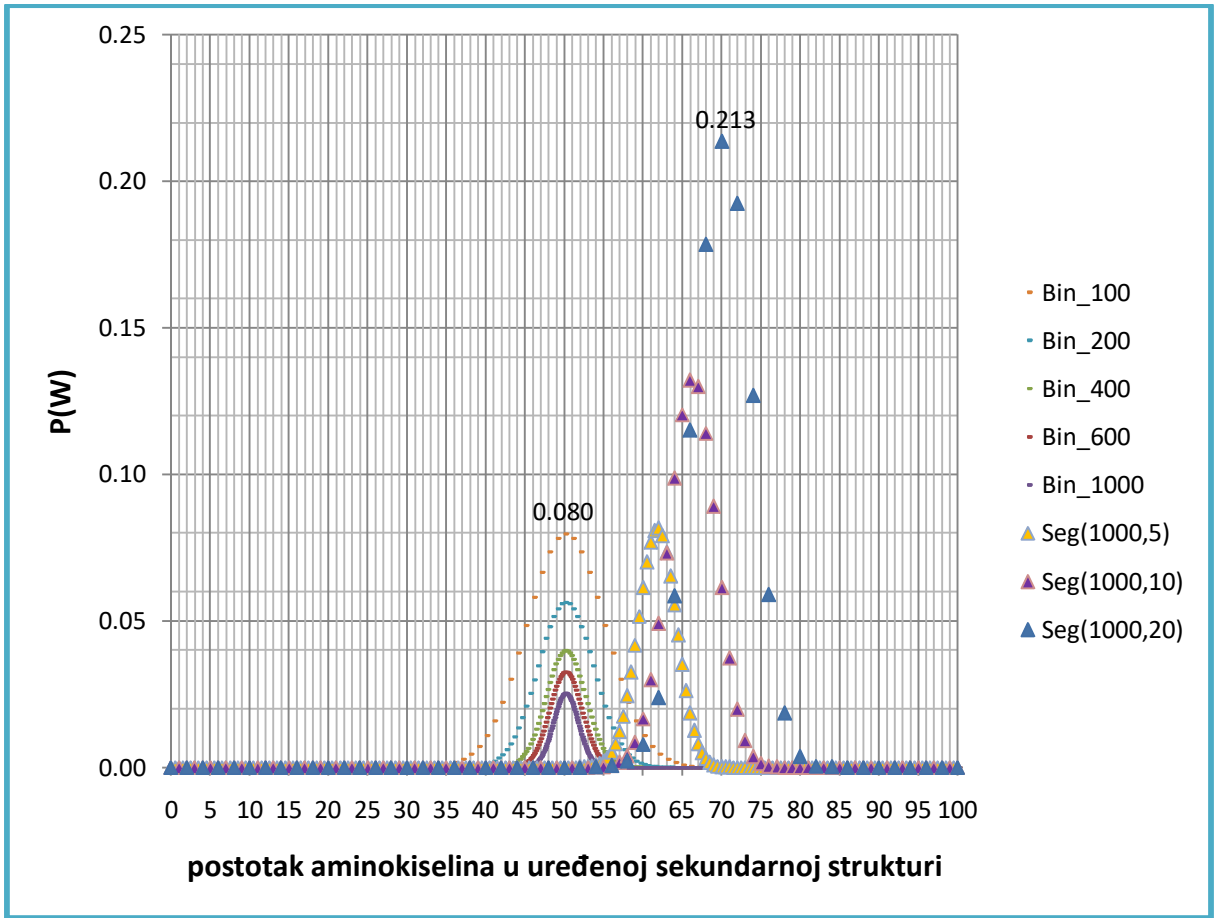
Slika 22. Ovisnost broja mogućih realizacija modelnih konformacija strukture binomnog (za slijedove duljina 100, 200, 400, 600 i 1000) i segmentnog nasumičnog modela (za slijed duljine 1000 aminokiselina i segmente duljina 5, 10, 20, 50 i 100) u ovisnosti o postotku aminokiselina u uređenoj sekundarnoj strukturi.

Vidi se razlika maksimuma broja realizacija modelnih konformacija strukture u slučaju binomnih modela (maksimumi se uvijek nalaze na 50% popunjenosti slijeda). Maksimumi broja realizacija za segmentni nasumični model pomiču se prema popunjenostima slijeda većim od 50%, a ovisni su o duljini slijeda i duljini segmenata, i uočava se asimetričnost. Kako bi se što zornije prikazala ova asimetričnost, na slici 23. prikazane su vrijednosti vjerojatnosti svih mogućih realizacija.

Kod binomnog modela vidi se da se sve vjerojatnosti maksimalnih brojeva realizacija modelnih struktura nalaze na popunjenosti 50%, i te vjerojatnosti opadaju s porastom duljine slijeda (N), dok gustoća broja realizacija raste s porastom N. Za duljinu slijeda $N = 100$ u intervalu popunjenosti $[45\% - 55\%]$ nalazi se 11 modelnih struktura s njihovim nasumičnim realizacijama, dok za duljinu slijeda 1000 u istom intervalu imamo 101 modelnu strukturu s njihovim nasumičnim realizacijama.

U slučaju vjerojatnosti različitih realizacija modelnih struktura u segmentnom nasumičnom modelu uočava se pomak vjerojatnosti koja odgovara maksimalnom broju realizacija prema postotku aminokiselina u uređenoj strukturi većem od 50. To utječe na broj modelnih struktura i na broj njihovih realizacija u pojedinom postotnom intervalu. Stoga, bilo bi korisno promotriti broj modelnih struktura i njihovih realizacija u intervalu $\pm 5\%$ oko maksimuma raspodjele.

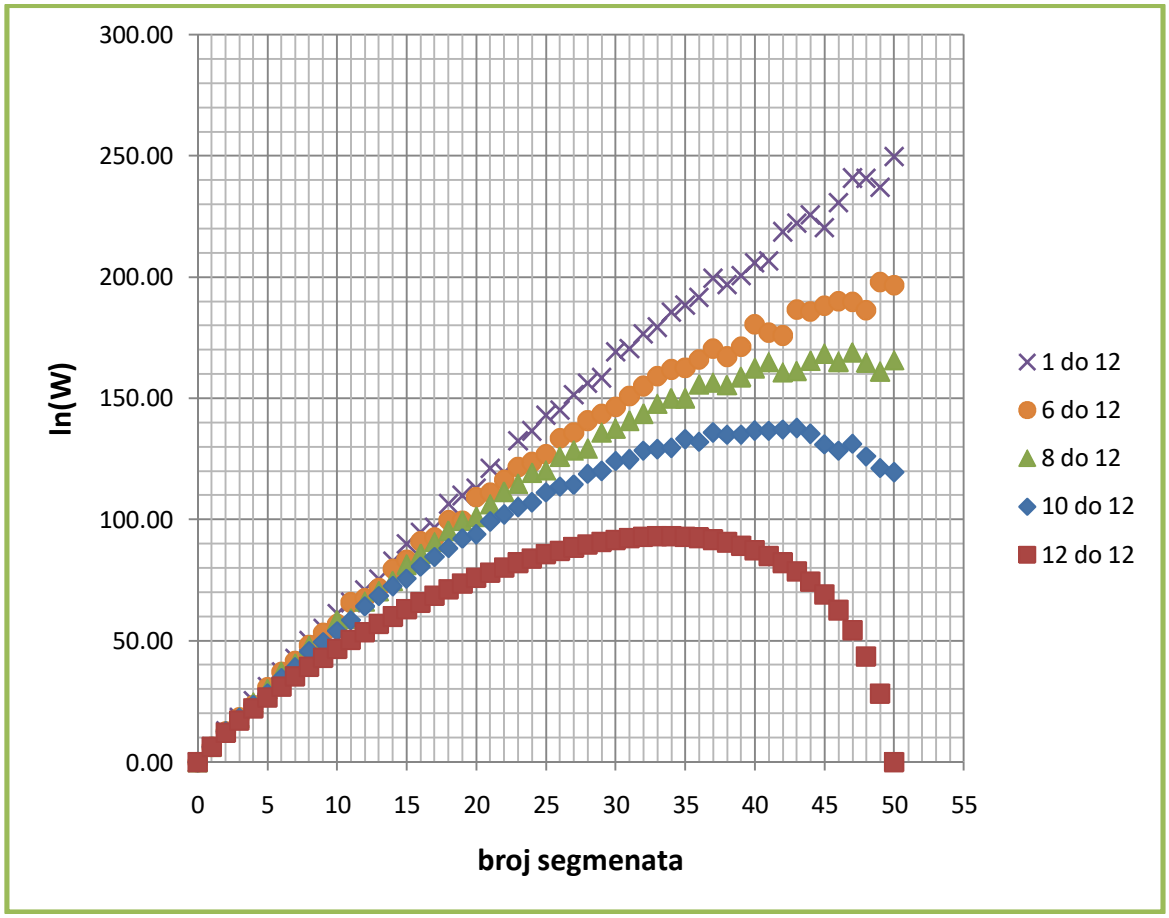
Istovremeno, gustoća stanja modelnih struktura u segmentnom nasumičnom modelu manji je u odnosu na binomni nasumični model (približno) onoliko puta koliko iznosi duljina segmenta u modelnim strukturama koje se promatraju i uspoređuju. U ovom primjeru taj je omjer 5 = 5/1, tj. segment duljine 5 u segmentnom nasumičnom modelu Seg(1000,5) podijeljen s duljinom segmenta u binomnom nasumičnom modelu koja je uvijek jednaka jednoj aminokiselini. (J. Batista; B. Lučić, rad u pripremi).



Slika 23. Vjerojatnosti stanja binomnog i segmentnog nasumičnog modela (za slijedove duljina 100, 200, 400, 600 i 1000) u ovisnosti o postotku aminokiselina u uređenoj sekundarnoj strukturi.

Napomena: za slučaj kada je duljina segmenta velika i usporediva s duljinom slijeda, raspodjele neće slijediti prethodno opisana pravila. Kako se duljina segmenta približava polovici duljine slijeda (imamo samo dva ili tri člana u raspodjeli i nije smisleno takav slučaj smatrati raspodjelom), vrijednost maksimuma može se značajno mijenjati za isti broj segmenata. Npr. za duljinu slijeda 1000, segmenti duljine 334 i 500 (maksimalno moguće ubaciti po dva segmenta) poprimaju značajno različite vrijednosti te je teško doći do kvalitetnih zaključaka.

Do sada su razmatrane raspodjele sa stalnim odnosom (duljina slijeda)/(duljina segmenta) ili promjenjive duljine segmenta uz stalnu duljinu slijeda. Razmotrimo sada promjenu broj mogućih realizacija za stalnu duljinu slijeda uz promjenjivu duljinu segmenta, ako duljine segmenata odabiremo nasumično. Neka je stalna duljina slijeda $N = 600$ u kojem odabiremo broj segmenata u rasponu od 1 do 50, čije se duljine mogu nasumično mijenjati u zadanom intervalu (u ovom slučaju najveća duljina TM segmenta je 12, jer je omjer duljine slijeda i maksimalno mogućeg broja TM segmenata jednak $600/50 = 12$). Na slici 24. prikazan je slučaj za nasumični odabir segmenata u intervalu od 1 do 12, kao i za kraće intervale uz stalnu gornju granicu. S obzirom da nasumično odabrani segmenti mogu biti kraći od 12, to se očekuje veće raspršenje vrijednosti u raspodjeli za slučaj (1 do 12), nego li za slučaj (6 do 12) ili kada je duljina segmenta točno 12 (12 do 12, slika 24). Svakako, u nasumičnim modelima mogu se dobiti i analizirati i raspodjele za veći broj segmenata, sve do potpunog popunjenja slijeda.

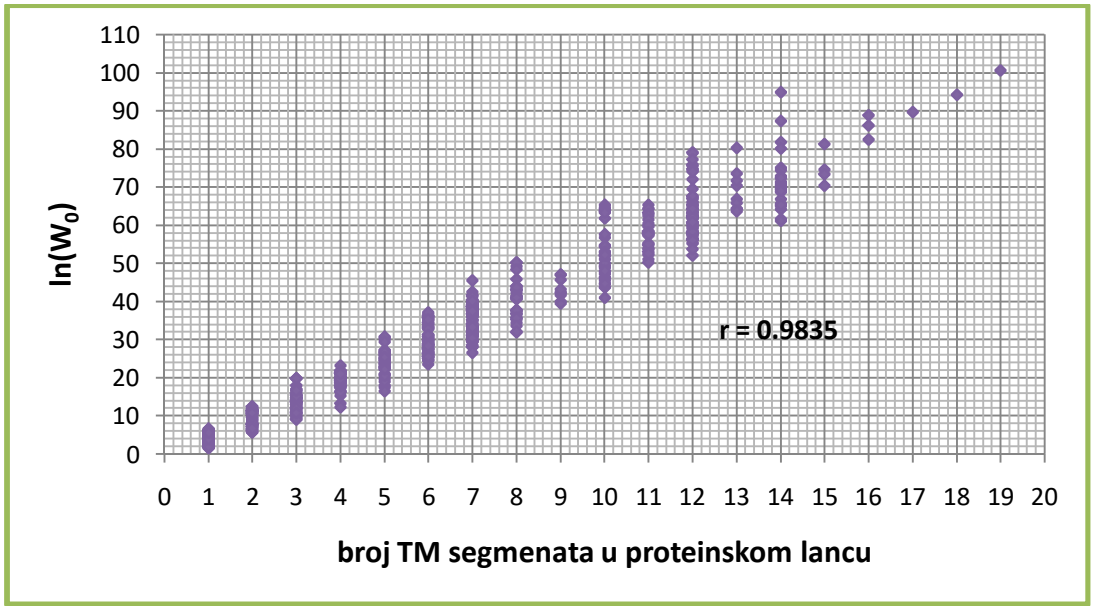


Slika 24. Vrijednost entropijskog koeficijenta ln(*W*) u segmentnom nasumičnom modelu u ovisnosti o broju segmenata i o varijabilnosti duljina segmenata.

Za realne proteinske lance varijabilnost je još i veća, jer istovremeno imamo varijabilnosti duljine segmenata, broja segmenata i duljine slijeda.

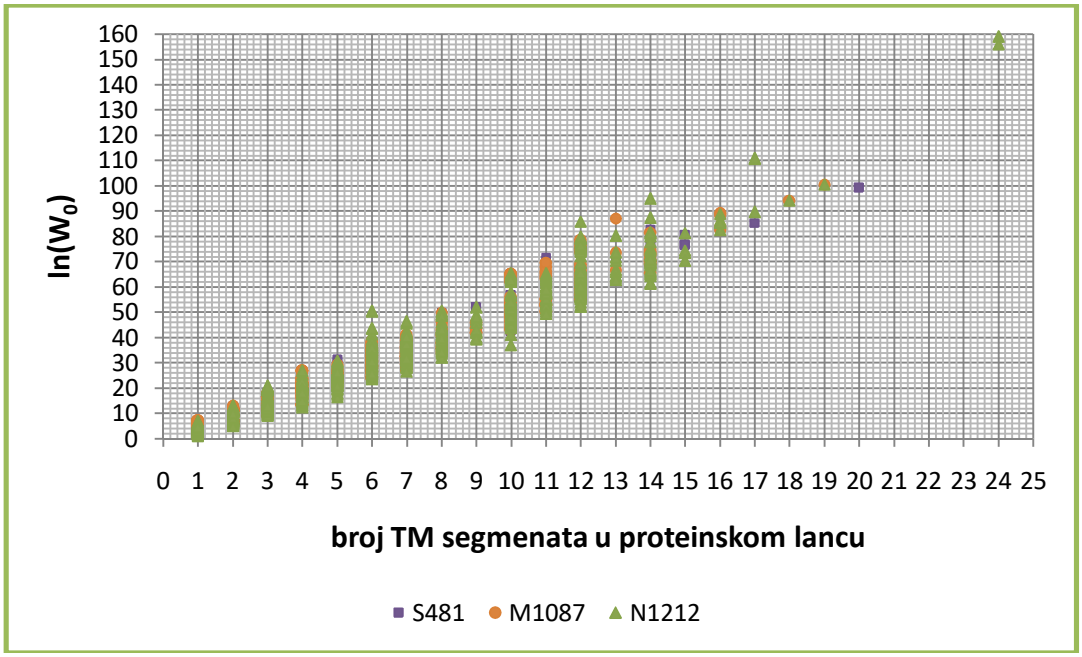
Stroga (značajna) veza između entropijskog koeficijenta u segmentnom nasumičnom modelu i broja TM segmenata u lancu daje za pravo da se u narednim razmatranjima umjesto korištenja entropijskog koeficijenta može koristi jednostavniji parametar, tj. broj TM segmenata u lancu. Dijagrami raspršenja za četiri opisana skupa u disertaciji: S481 (poglavlje 2.4.2.), M1087 (poglavlje 2.4.3.), N1212 i N907 (poglavlje 3.3.1.), dani su na slikama 25. i 26. Ovisnost broja realizacija modelnih struktura prema segmentnom nasumičnom modelu o broju TM segmenata u lancima prikazana je na slici 25 za stvarni skup izdvojen u disertaciji koji se sastoji od 907 proteinskih struktura (riješenih s rezolucijama do 3.5Å) definiranog u poglavlju 3.3 kao početni skup N907.(J. Batista; B. Lučić, rad u pripremi).

Točke sa vertikalnim pomakom za isti broj TM segmenata odgovaraju lancima različitih duljina, i prema segmentnom nasumičnom modelu veće vrijednosti entropijskog koeficijenta odgovaraju duljim proteinskim slijedovima.



Slika 25. Ovisnost logaritma broja realizacija modelne strukture prema segmentnom nasumičnom modelu o broju TM segmenata u proteinskom lancu za skup N907.

Uočava se velika zastupljenost proteina sa 8, 10, 12 i 14 TM segmenata (velika razvučenost na vertikalnim pravcima za ove vrijednostina osi x) koji se značajnije razlikuju u duljinama. Pod-skupine lanaca koji imaju 1, 2, 3, 6 i 11 TM segmenata po lancu pokazuju kompaktnije ponašanje, što ukazuje na to da su ti lanci ujednačenijih duljina.



Slika 26. Usporedba ovisnosti logaritama broja realizacija modelne strukture (entropijskih koeficijenata) prema segmentnom nasumičnom modelu o broju TM segmenata u proteinskom lancu za početne skupove M1087, S481 i N1212.

Na slici 26. dana je usporedba analogna onoj na slici 25 za početne skupove korištene u u disertaciji. Vidi se značajna korelacija između entropijskih koeficijenata i broja TM segmenata u proteinskim lancima (korelacijski koeficijenti veći > 0.98). Vrijednosti korelacijskih koeficijenata za skupove su: $r_{\mathrm{M1087}} = 0.9843$; $r_{\mathrm{S481}} = 0.9874$; $r_{\mathrm{N1212}} = 0.9804$.

3.1.6. Algoritmi 1 i 2 – algoritmi slični algoritmima Hobohm 1 i 2

Analizom primjena algoritama iz literature za izbor reprezentativnih skupova membranskih proteina u sklopu izrade disertacije uočeno je kako se njima 'po definiciji' nastoji dobiti maksimalni broj proteinskih lanaca bez razmatranja osobina ili kvalitete strukture (npr. sekundarne strukture lanca) [24,25,39,40,41,46].

U nastojanju dobivanja što je moguće sadržajnijeg reprezentativnog skupa membranskih proteina niske međusobne sličnosti/identičnosti, u koji će se birati proteinski slijedovi sa što složenijom strukturom (tj. da uravnotežena vrijednost parametra $Q_{2,rnd}$ bude što bliža 50%, i što veći broj mogućih realizacija modelne strukture W_i), koji će biti i najkvalitetniji i najzahtjevniji u modeliranju strukture, krenulo se najprije od standardnih algoritama "Hobohm 1" i "Hobohm 2".

Uvedene se i korištene kao osnovne varijable u algoritmima podaci karakteristični za pojedini proteinski lanac i njegovu strukturu. Ti su podaci dostupni na mrežnim stranicama baza OPM ili PDB poput: pdb identifikatora (PDB kod), proteinskog slijeda i njegove duljine, rezolucije, eksperimentalna metode kojom je određena struktura ili vrste proteina. Drugi dio varijabli potrebnih u radu algoritama za izbor raprezentativnih skupova proteina prvi put je definiran i uveden u uporabu u disertaciji u nastojanju kvantificiranja složenosti sekundarne strukture proteina. Te se varijable računaju na temelju informacija iz primarne i sekundarne strukture membranskih proteina, a najvažnije su: (1) broj trasmembranskih segmenata, (2) koeficijent $Q_{2,rnd}$ koji kvantificira najvjerojatniju nasumičnu točnost balansiranoga modela s dva stanja, (3) binomni koeficijent koji broji sve moguće nasumične realizacije modelne strukture proteina (na temelju broja aminokiselina u dijelu proteina koji je u membrani), te (4) broja svih mogućih nasumičnih realizacije modelne strukture proteina prema segmentnom modelu (u kojem se elementi strukture promatraju grupirani u segmente).

U samim je temeljima rada algoritama iz literature njihova velika ovisnost o redoslijedu lanaca u listi početnoga skupa. Međutim, nisu susatvno razrađeni niti analizirani kriteriji izbora optimalnog redoslijeda lanaca (tj. sortiranja/redanja po nekoj izabranoj varijabli/svojstvu) koji će dati najveći konačni reprezentativni skup niske međusobne sličnosti (niže od unaprijed zadanoga praga). U analizama se uočava velika varijabilnost konačnog skupa proteinskih slijedova u ovisnosti o sortiranjima po pojedinoj varijabli. Kako bi se izbjegli navedeni problemi, u disertaciji je analizirana, između ostalog, i mogućnost proizvoljnog nasumičnog (engl. *random*) odabira poretka u onim slučajevima u kojima se prema odabranom kriteriju/varijabli za sortiranje javlja više proteinskih lanaca s istim ishodom (vrijednošću). Kao druga mogućnost u rješavanju toga problema nametalo se uvođenje više kriterija za sortiranje (s tim da je jedan kriterij primarni, drugi, sekundarni, itd.), što usložnjava algoritam i shodno tome produljuje njegovo izvođenje. U nastojanju popravki postojećih algoritama nastojalo se:

- dobiti maksimalna vrijednost za cjelokupni skup uz pojedini kriterij (npr. ukupni broj transmembranskih segmenata u skupu ili ukupni $Q_{2,rnd}$ u skupu), što ne vodi nužno i najvećem broju proteinskih lanaca,
- dobiti što je moguće brži algoritam.

Zajednički dio programskoga koda u algoritmima 1 i 2 omogućuje da se, uz najvažnije definirane osobine (gore spomenute varijable), prilikom unosa (učitavanja) podataka proteinskom lancu pridijeli redni broj, ili da se početni skup lanaca razvrsta u podskupove po raznim odabranim kriterijima poput (1) eksperimentalne metode kojom je određena struktura, (2) strukturne vrste proteina (npr. *bitopic* ili *polytopic*) te da se izdvoje samo (3) proteini čija je struktura riješena s rezolucijom ispod unaprijed definiranoga praga/granice. Potom se učitavaju podaci u obliku uređene ulazne liste podataka za odabrane lance početnoga skupa po prethodnim kriterijima, te matrica sličnosti (identičnosti) između proteinskih lanaca s definiranim pragom sličnosti.

U cilju povećanja brzine rada algoritma matrica sličnosti (identičnosti) sa sličnostima izraženim postotcima u rasponu od 0 do 100% pretvorena je, uporabom praga sličnosti, u binarnu matricu. U takvoj matrici vrijednost 1 za neki par lanaca znači sličnost koja je jednaka ili veća od praga sličnosti, a vrijednost 0 znači da je sličnost između para proteina manja od praga sličnosti (tj. među njima nema zalihosti). Spremanje binarne matrice sličnosti, koja je kvadratna matrica, zahtijeva manje memorijskoga prostora i smanjuje količinu spremljene ili dohvaćene informacije u svakom koraku izvođenja algoritma. Algoritmi razvijeni u disertaciji rade sa simetričnom matricom sličnosti, iako ona može biti i nesimetrična, kakva je u pravilu u algoritmu UniqueProt [46]. Naime, za dva lanca A i B, različite dužine, relacija „*lanac B sličan je lancu A*“ (npr ako je ta sličnost malo iznad definiranoga praga) ne uvjetuje nužno i postojanje obratne realacije koja bi glasila „*lanac A sličan je lancu B*“ (ta sličnost može biti malo ispod definiranoga praga sličnosti. U slučaju promjene praga sličnosti, postupak učitavanja binarne matrice mora se provesti ponovno. Za matricu dimenzije 1000×1000 (trajanje učitavanja matrice iznosi između $10s - 20s$) ubrzanje je značajno u odnosu na učitavanje matrice sa decimalnim (float) vrijednostima.

Nakon učitavanja matrice sličnosti bira se kriterij (varijabla) po kojoj se provodi uređivanje (redanje, tj. sortiranje) proteina u početnome skupu. Pritom se koriste dva standardna načina, slična onima koji se koriste u postojećim algoritmi Hobohm 1 i Hobohm 2 [24,25] te su, analogno tome, odgovarajući algoritmi razvijeni u disertaciji nazvani Algoritam 1 i Algoritam 2. Osnovna razlika ogleda se u tome da se u Algoritmu 1 korisnički odabrani (po nekoj varijabli, ili po više njih) proteinski lanac ostavljanja, dok se u Algoritmu 2 on odbacuje.

Algoritam 1

Programski kod za Algoritam 1 napisao sam po opisu prvoga algoritma (nazvan kasnije u literaturi kao Hobohm 1) za izbor reprezentativnog skupa proteina niske međusobne sličnosti, a prema opisu danom u radu autora Hobohm i dr. iz 1992 godine [24]. Opis rada algoritma razložen je u sljedećim točkama:

(i) Za svaki proteinski lanac u ulaznome skupu (listi) zbroje se svi lanci koji su mu slični više nego je postavljeni prag sličnosti. Tako se za svaki protein dobiva informacija o veličini podskupa lanaca koji su mu previše slični (tj. iznad definiranog praga 'prihvatljive' sličnosti).
(ii) Ulazni skup (N) lanaca uređuje se (sortira) prema jednom ili više kriterija (od kojih je jedan primarni, pa potom slijedi sekundarni, tercijarni, ...) koje zadaje korisnik (npr. veličina podskupa proteinskih lanaca sličnih pojedinom lanacu, rezolucija kojom je određena struktura, duljina lanca, broj TM segmenata).
(iii) U konačni skup reprezentativnih proteinskih lanaca bez zalihosti dodaje se prvi proteinski lanac po redu u uređenom ulaznom skupu (u početnom skupu ostaje $N - 1$ lanaca za analizu),
(iv) Identificira se (izdvoji) podskup (n_1) lanaca koji su (prvom) proteinskom lancu izabranom u reprezentativni skup pod (iii) slični više nego je dopušteni prag sličnosti. Potom se svi takvi proteinski lanci izdvoje (odbace) iz ulaznoga skupa, u kojem za analizu preostaje $(N - 1 - n_1)$ lanaca.
(v) ponovno se provede postupak opisan pod (i) i potom se ponavlja analogan postupak za sljedeći proteinski lanacna koji algoritam nailazi u ulaznom skupu, a sve se ponavlja sve dok se ne iscrpe svi proteini u ulaznome skupu.

Analiziranjem odabira proteinskog lanca u pojedinoj iteraciji uočilo se da izbor lanca po pojedinom kriteriju nije nužno jednoznačan nego se, uz neke izabrane kriterije uređivanja (sortiranja) ulaznoga skupa (npr. broj TM segmenata), može pojaviti više proteinskih lanaca s

istom vrijednošću. Takva je višestrukost izraženija ako je kao kriterij uređivanja izabran broj TM segmenata, jer svi lanci sa TM segmentima imaju između 1 i 24 TM segmenta (zači, sva različitost u tom smislu može se opisati s 24 različita broja), nego ako je kao kriterij izabrana rezolucija koja je iskazana na dvije decimale i postoji više od 24 različitih vrijednosti rezulucije u PDB [20]. Najboljim se pokazalo ako se kao prvi kriterij koristi silazno uređivanje (sortiranje) ulaznog skupa po broju TM segmenata (prvi lanac ima najviše TM segmenata), te ako se kao sekundarni kriterij uređivanja koristi broj lanaca sličnih odabranom lancu iz ulaznoga skupa i to uzlazno (prvi je lanac koji ima najmanje sebi sličnih lanaca).

Ovakvim načinom u reprezentativni skup dodaju se (biraju) proteinski lanci koji, uz zadani broj njemu sličnih lanaca, ima najveći broj TM segmenata. U slučaju da su oba kriterija jednaka za više od jednog lanca, kako bi se izbjegla ovisnost o ulaznom (abecednom) uređivanju, proteini koji imaju jednaki broj TM segmenata poredaju se nasumično na samom početku. Uvođenje ovakvog nasumičnog odabira lanaca davalo je bolje rezultate, s tim što se za veće skupove ulaznih lanaca produži vrijeme rada algoritma.

Zaključak: Uspješnost Algoritma 1 ovisi o početnim (ili tamo gdje se primjenjuje, o kasnijim) uređivanjima (sortiranjima), a rezultati mogu biti značajno različiti pri izboru različitih kriterija uređivanja. Nadalje, vrijeme rada algoritma ovisi o tome uvodi li se nasumični odabir poretka za proteina koji imaju istu vrijednost kriterija u petlji (u kojoj se zadaje veći broj iteracija).

U tablici 17. dana je analiza skupa S392 Algoritmom 1 na razini sličnosti između proteinskih lanaca od 30%, i to za različite kriterije uređivanja (sortiranja) koje se obavlja prije petlje u kojoj se preračunava broj susjeda za svaki protein. Pritom, unutar petlje ne rade se dodatna sortiranja.

Tablica 17. Početni rezultati dobiveni Algoritmom 1 u analizi skupa S392 (sličnost 30%).

kriteriji uređivanja	sortiranja u ulaznim linijama (klasteri proteina definirani ulaznim listama, a ne gleda se broj susjeda u klasteru)								
PDB kod	uzl.	sil.							
duljina slijeda			min	max					
broj TM segmenata					min	max			
rezolucija							min	max	
nasumično									da*
broj prot. lanaca	94	103	97	101	99	98	98	96	102
broj TM segmenata	442	498	446	477	450	468	444	473	479
broj AK	28066	31290	27304	31490	28888	28087	27588	29746	29615

'min' = ulazni skup uređen (složen) počevši od najmanje vrijednosti kriterija
'max' = ulazni skup uređen (složen) počevši od najveće vrijednosti kriterija
'uzl.' = ulazni skup uređen (složen) uzlazno alfabetski
'sil.' = ulazni skup uređen (složen) silazno alfabetski
'da' = korištena nasumična preraspodjela proteinskih lanaca s istim prethodnim kriterijima pri odabiru lanca koji se ostavlja: znak ' * '= izabran najbolji rezultat od 30 iteracija

Dobiveni rezultati ne razlikuju se značajno od vrijednosti dobivenih za reprezentativni skup S101 koji ima 101 proteinski lanac s 483 transmembranska segmenta i 30144 aminokiseline. To je glavni reprezentativni skup integralnih membranskih proteina alfa vrste međusobne sličnosti ispod 30% izabran algoritmom Hobohm 2 [25] iz početnog skupa S392 a objavljen u radu 'grupe Sydney' [41]. Ipak Algoritmom 1 dobiven je u jednoj kombinaciji (drugi skup po redu u tablici 17) i nešto bolji skup gledano i po broju TM segmenata (498) i po broju lanaca (103).

Ako se primjenjuje samo kriterij nasumičnog izbora lanaca u tablici 17. (označeno s "da*" u krajnjem desnom stupcu), rezultat se može značajno mijenjati po pojedinoj iteraciji

(kojih je ukupno 30). Zamislimo da se nastoji proći sve moguće kombinacije poredaka lanaca unutar svakoga klastera proteina koji su međusobno previše slični, i ako bismo kombinirali sve moguće nasumične iteracije, jedna od njih dala bi optimalni maksimum npr. prema broju lanaca. No, zbog velikog broja svih mogućih kombinacija za ovaj skup to nije moguće uraditi u realnom vremenu (broj svih mogućih kombinacija bio bi veći od 10^{39}).

Algoritam 2

Programski kod Algoritma 2 napisan je prema opisu algoritma Hobohm 2 [25]. Ovaj algoritam ima iste korake kao prethodno opisani Algoritam 1 s tom razlikom da se:

(a) Uvijek izbacuje i iz osnovnoga skupa i iz daljnjih analiza onaj proteinski lanac koji ima najveći podskup (klaster) sebi previše sličnih proteina.

(b) Ovim načinom rada, u svakoj iteraciji izbacuje se točno jedan proteinski lanac, stoga ovaj algoritam treba uraditi točno onoliko iteracija koliko ima proteinskih lanaca koji imaju barem jedan lanac u skupu s kojim su slični više nego je dopuštena razina (prag) sličnosti.

(c) Konačni reprezentativni skup lanaca niske međusobne sličnosti čine lanci iz preostalog podskupa u kojemu niti jedan proteinski lanac nema niti jednoga lanca koji mu je sličan više nego je definirani prag maksimalne sličnosti.

(d) Za velike početne skupove proteinskih lanaca, ovaj algoritma radit će jako sporo. To će se dodatno usporiti ukoliko se dodaju petlje ukojima se provode iteracije s ciljem pronalaženja optimalnog poretka i optimalnoga kriterija za uređivanje (sortiranje) skupa.

S obzirom da je algoritam Hobohm 2 davao najbolje rezultate u smislu najvećeg broja proteinskih lanaca u konačnom reprezentativnom skupu lanaca niske međusobne sličnosti, bio je najčešće korišten algoritam za te svrhe prema literaturnim izvorima i prema učestalosti citiranja u literaturi. Bitno je spomenuti da je algoritam Hobohm 2 naveden je i kao standard u PDB [20] za izdvajanje skupa lanaca niske međusobne sličnosti. Kvaliteta rezultata koji se dobivaju Algoritmima 1 i 2 bit će se prikazati u poglavlju Rezultati.

3.1.7. Algoritam 3 – algoritam temeljen na broju zajedničkih susjeda

U analizama rezultata na raznim skupovima proteinskih lanaca dobivenih algoritmima 1 i 2, kao i u rezultatima koje su dobili drugi autori drugim postupcima [39,40,41], uočene su mogućnosti poboljšanja algoritama za izbor reprezentativnih skupova lanaca. Na temelju toga, i analizom i promišljanjem problema izbora reprezentativnog skupa proteina, razvijen je algoritam nazvan Algoritam 3, koji predstavlja originalni doprinos.

Opis algoritma 3

A) Učitavanje i priprema podataka

Najprije se učitavaju informacije o strukturnim svojstavima primarne i sekundarne strukture iskazana brojčano. Potom se odabire vrijednost praga sličnosti, tj. razina najveće dopuštene (prihvatljive) sličnosti između proteinskih slijedova u reprezentativnom skupu, i podaci o sličnosti svih parova lanaca organiziraju se u obliku matrice sličnosti proteinskih lanaca (dimenzije $N \cdot N$). Radi ubrzanja rada algoritma vrši se simetrizacija matrice sličnosti i njezina dskretizacija, gdje se za sličnosti veće od praga stavlja vrijednost 1, a za sličnosti koje su manje od praga, vrijednost u matrici postaje jednaka 0.

Kako bismo ilustrirali problem koji se može pojaviti pri kvantificiranju sličnosti a koji se treba riješiti u osmišljavanju algoritama za redukciju sličnosti među proteinskim lancima, promotrimo izlazne vrijednosti identičnosti i sličnosti koje se dobiju primjenom programa

EMBOSS_needle [44] za lanace1fjk_A i 2x2v_A. U tablici 18. vidimo da vrijednosti ovise o tome koji je lanac postavljen na ulazu kao prvi a koji kao drugi.

Tablica 18. Primjer asimetričnih vrijednosti identičnosti i sličnosti ovisno o poretku lanaca.

lanac 1	lanac 2	identičnost(%)	sličnost(%)
1fjk_A	2x2v_A	16.7	29.2
2x2v_A	1fjk_A	17.8	32.9

Pretpostavimo li tako da je granica sličnosti 30%, onda bi se u jednom poretku dogodilo da su oba proteina prihvatljive sličnosti, i niti jedan ne bi trebao biti izuzet iz konačnog reprezentativnog skupa. Međutim, promatrani u obrnutom poretku, ova dva lanca imaju sličnost veću od 30% i jedan od dva lanca morao bi biti izuzet iz konačnog reprezentativnog skupa. Ako je jedna od ove dvije vrijednosti iznad korisnički definirane vrijednosti, s obzirom na provođenje simetrizacije matrice sličnosti u Algoritmu 3, smatra se da su lanci međusobno previše slični bez obzira na poredak. Ovakav postupak simetrizacije zapravo postrožava postupak izbora reprezentativnog skupa proteinskih lanaca međusobne sličnosti ispod zadanog praga sličnosti.

Osim matrice sličnosti u radu algoritma potrebno je uzeti u obzir i svojstva primarne i sekundarne strukture proteinskih lanaca, te se matrici sličnosti dodaju stupci u kojima su podaci koji se izražavaju cijelim brojevima, poput rednog broja lanca, broja TM segmenata, duljine slijeda ili rezolucije. U slučaju korištenja rezolucije, vrijednost iz PDB množi se sa 100 kako bi se dobile cjelobrojne vrijednosti. Na taj se način u proširenoj matrici sličnosti zapravo definira jedinstveni zapis za svaki proteinski lanac, koji sadrži sve potrebne informacije koje algoritam kasnije koristi u svome radu.

B) Redukcija ulaznoga skupa proteinskih lanaca (osnovno pročišćavanje)

Ulazna se matrica sličnosti najprije reducira za sve one lance koji su jedinstveni (odnosno koji nemaju sličnost iznad praga niti s jednim drugim lancem iz skupa), i kao takvi odmah su izabrani u konačni reprezentativni skup.

Potom Algoritam 3 (u početnoj analizi matrice sličnosti) identificira neke jednostavne a korisne slučajeve, te tako skraćuje i unapređuje izbor reprezentativnog skupa. Naime, ako su dva redka u matrici identična, to znači da dva proteinska lanca kojima ti redci odgovaraju imaju *identičan podskup lanaca koji su im slični više od praga dopuštene (prihvatljive) sličnosti*. Takvi lanci smatraju se identičnima, i jedan od njih mora se izuzeti iz daljnje analize (zadržat će se od ta dva lanca onaj koji je povoljniji prema vrijednosti odabranoga kriterija koji se nastoji maksimizirati u konačnom reprezentativnom skupu, npr. broj TM segmenata ili rezolucija). Ovakav način pročišćavanja ulaznoga skupa i eliminacija suvišnih proteinskih lanaca može se provesti bez obzira na sve daljnje odluke u radu algoritma, na koje taj postupak neće utjecati. Kako bismo to i potkrijepili, promotrimo sljedeći primjer: pretpostavimo da se u početnom skupu nalazi veći broj od $N_p (N_p \geq 2)$ lanaca koji su međusobno slični iznad praga sličnosti, i da svi ti lanci zajedno imaju isti podskup lanaca koji su im slični više od praga prihvatljive sličnosti (taj podskup nazovimo kraće kao *podskup lanaca istih susjeda*). Ako bilo koji lanac iz takvog podskupa zadržimo, sve ostale moramo izbaciti. Međutim, ako zadržimo i bilo koji lanac l_i koji nije u tom podskupu lanaca istih susjeda, a svi lanci iz tog podskupa imaju upravo taj lanac l_i kao zajedničkog susjeda, nužno se taj cijeli podskup lanaca mora izbaciti. Na temelju toga zaključujemo da reduciranje matrice sličnosti na način da se od svakog pronađenog *podskupa lanaca istih susjeda* zadrži samo jedan lanac optimalnih osobina i strukturnih svojstva (onih koje želimo optimirati i na konačnom skupu), ne utječe negativno na krajnji rezultat (tj. kvalitetu reprezentativnog skupa) nego je u odnosu na krajnji rezultat neutralno. U radu Algoritma 3

optimiran je konačni skup na način da se nastoji imati proteinske lance s najvećim brojem TM segmenata i da se pritom (kao sekundarni kriterij) nastoji izabrati lanci maksimalne kompleksnosti (a samo u početnoj ulaznoj listi lanci su složeni počevši od bolje rezolucije). Stoga, od svakog *podskupa lanaca istih susjeda* izabiran je u Algoritmu 3 kao predstavnik onaj lanac koji ima najviše transmembranskih segmenata, i on će jedini ući u daljnje analize.

Nakon toga dijela algoritma za analizirati ostaju samo oni lanci koji i dalje imaju neke zajedničke susjede ali međusobno nisu slični više od praga sličnosti, i oni su kandidati za jedinstvene proteinske lance za reprezentativni skup (i kao takvi idu u sljedeći glavni dio algoritma).

C) Izbor jedinstvenih lanaca u reprezentativni skup niske međusobne sličnosti

U glavnoj petlji algoritma uzima se pročišćena lista proteinskih lanaca iz prethodnoga dijela i pročićena (reducirana) matrica sličnosti. Konačni rezultat koji se treba postići radom algoritma dodatna je (što je moguće štedljivija) redukcija suvišnih (ne-jedinstvenih) lanaca i redukcija matrice sličnosti, sve dok reducirana matrica u nekom koraku ne bude imala izvan dijagonale sve vrijednosti 0 (a na dijagonali 1). U nastavku je opis dijelova Algoritma 3.

i. Sljedeći je korak učitavanje i zbrajanje svih susjeda (tj. onih lanaca koji su mu slični više nego je definirani prag sličnosti) iz matrice sličnosti za svaki pojedini lanac. Potom se za svaki proteinski lanac izračuna nova varijabla koja je *omjer broja TM segmenata koji taj lanac ima i broja susjednih lanaca* (izračunat u prethodnom koraku), i po toj novoj varijabli (gledano silazno) poslože se svi lanci u skupu. Tako da se u posloženoj (uređenoj) listi lanaca kao prvi nađe lanac s najvećom vrijednošću ove varijable (omjera). Može se definirati i neki drugi parametar i po njemu provesti uređivanje (sortiranje) skupa lanaca, ali ovaj je parametar dao najbolje rezultate na analiziranim skupovima u disertaciji. Radi ilustracije promotrimo primjer usporedbe dva lanca, od kojih prvi ima dva TM segmenta i dva susjeda (po sličnosti koja je iznad definiranoga praga), a drugi ima dva TM segmenta i četiri susjeda. Ovim postupkom zadržava se lanac s dva susjeda, jer se njegovim zadržavanjem izbacuje manji broj drugih lanaca.
ii. Nakon odabira prvog lanca (koji ima najveći omjer broja TM segmenata i broja susjednih lanaca), gledaju se u programskoj petlji svi drugi lanci i razmatra kakva je razlika u podskupovima susjeda, odnosno gleda se presjek podskupova susjeda drugih lanaca s odabranim. Ako je razlika podskupova susjeda takva da je to prazan skup (što znači da ne treba dodatno izbacivati niti jedan lanac) i ukoliko taj lanac nije sličan (više od praga sličnosti) odabranom lancu, tada se zadržava i taj lanac. To znači da ako odaberemo neki lanac za zadržavanje i ako drugi lanac ima iste susjede kao i on ali nisu međusobno slični, tada se u ulaznom skupu zadržavanjem ovog drugog lanca neće izbaciti niti jedan drugi lanac. Ako ova razlika daje određeni broj novih lanaca koji bi se trebali izbaciti, gleda se omjer broja segmenata u tom lancu i zbroja TM segmenata drugih lanaca koji su susjedni ovom lancu, te se odlučuje do koje razine tog omjera treba ići. U Algoritmu 3 korištena je vrijednost ovog omjera od 0 (tj. > 0) do 2.0 gdje je uočeno da više vrijednosti nisu davale bolje rezultate u izboru reprezentativnog skupa. Zatim se petlja prekida i ide na novu iteraciju pri čemu se uzima novi lanac kao kandidat za zadržavanje. Iteracije nastavljamo sve dok ne dobijemo konačni skup u kojem više nema novih lanaca za analizu i odabir, tj. matrica koraka postane ortogonalna.

Zaključak: Ideja o razmatranju zajedničkih susjeda pri odlučivanju o zadržavanju proteinskog lanca u konačnom reprezentativnom skupu potpuno je nova, i nema analoga niti u jednom drugom algoritmu.

Uvođenjem omjera broja transmembranskih segmenata i broja susjeda dobiva se novi kriterij na osnovu kojega se donosi odluka o zadržavanju (ili izbacivanju) proteinskog lanca.

Pritom se nastoji zadržavanjem novog proteinskog lanca izbaciti što je moguće manji broj drugih proteinskih lanaca i TM segmenata, i na taj se način omogućuje globalna optimizacija odabranoga svojstva u reprezentativnom skupu. Ako se za odabrani skup optimira (maksimizira) npr. broj TM segmenata (odabrana varijabla), to nužno ne znači da će se dobiti i skup s najvećim brojem lanaca, ili pak s najvećim ukupnim brojem aminokiselina. Svakako, algoritam se može izmijeniti tako da se gleda i optimira ne samo jedna nego i neka druga varijabla, ali bi se s time problem optimizacije usložnio jer bi jako porastao broj iteracija.

3.2. Rezultati dobiveni primjenom Algoritama 1, 2 i 3 na skupove drugih autora

Kako bi se provela međusobna usporedba dobivenih algoritama, kao i s postojećim algoritmima Hobohm 1, Hobohm 2 [24.41] i UniqueProt [44] analizirani su sljedeći skupovi:

- S481 i S392 na granicama od $20\% - 35\%$ (usporedba prvenstveno u odnosu na rezultate 'grupe Sydney' i njhove primjene Hobohm 2 algoritma),
- radi usporedbe s algoritmom UniqueProt analizirani su originalni skupovi 'grupe München' M190 i M1087, kao i skupovi S148 (lanci tog skupa su podskup lanaca skupa S481) i N189 i N263. Skupovi N189 i N263 podskupovi su originalnog skupa N1212 dobivenog izdvajanjem lanaca iz baze OPM.

Primjenom Algoritama 1 i 2, kao i originalnog Algoritma 3, analizirali su se i uspoređivali izabrani reprezentativni skupovi s obzirom na:

a) ukupni broj lanaca,
b) ukupni broj TM segmenata,
c) prosječni prametar $Q_{2,\mathrm{rnd}}$,
d) ukupnu složenost izabranog skupa prema segmentnom modelu (model u kojem je minimalni razmak TM segmenata jednak nuli, označen sa S_0).

Ova usporedba kvalitete izabranih reprezentativnih skupova proteina daje informaciju o kvaliteti i efikasnosti novih i postojećih algoritama. Svi algoritmi polaze u radu od istih početnih skupova, zaključujemo kako su najkvalitetniji oni algoritmi koji u konačni reprezentativni skup izabiru proteinske lance najsloženije (najzahtjevnije) strukture. Dodatno, analizira se i ukupni izabrani broj proteinskih lanaca, ukupni (i prosječni) broj TM segmenata u skupovima, te ukupni brojevi pojedinih pod-skupina proteinskih lanaca s određenim brojem TM segmenata.

3.2.1. Rezultati dobiveni na skupu M190

Skup M190 reprezentativni je skup "Grupe Munchen" izabran iz početnog skupa od 1101 proteinskog lanca [42]. Početni skup od 1101 lanca nije ponuđen u radu, niti ga je bilo moguće izdvojiti iz baza OPM [29] i PDB [20] zbog stalnih pročišćavanja struktura i promjena u tim bazama. Rezultati analize reprezentativnog skupa M190 s pomoću različitih algoritama dani su ukratko u tablici 19. Početni skup M190 sadrži lance koji su izabrani prvom inačicom algoritma UniqueProt [42] objavljenom u literaturi za membranske proteine. Proteinski lanci izabrani su tako da je njihova međusobna identičnost manja od 20%, a analiza identičnosti provedena je algoritmom UniqueProt [44], što je opisano u pod-poglavlju 2.3.2.

Tablica 19. Analiza reprezentativnog skupa M190 dobivenog algoritmom UniqueProt [42] i algoritmima razvijenim u disertaciji.[a]

parametri / algoritam	*lanaca*	AK_{uk}	TM_{uk}	$Q_{2,rnd,sr}$	Bin_{uk}	$S_{0,uk}$	TM_{sr}	$S_{0,sr}$	$Q_{2,rnd}$
UniqueProt-2	161	32294	506	0.604	16672	2367	3.14	14.70	0.56
Algoritam 1	161	31659	503	0.601	16471	2348	3.12	14.58	0.55
Algoritam 2	161	32095	509	0.601	16747	2381	3.16	14.79	0.55
Algoritam 3	161	32095	509	0.601	16747	2381	3.16	14.79	0.55
Algoritam 1n	161	32095	509	0.601	16747	2381	3.16	14.79	0.55
Algoritam 2n	160	31944	500	0.604	16456	2339	3.13	14.62	0.56

[a]objašnjenja kratica: *lanaca* – ukupan broj lanaca u reprezentativnom skupu (RS); AK_{uk} – ukupni broj aminokiselina u RS; TM_{uk} – ukupni broj TM segmenata u RS; $Q_{2,rnd,sr}$ – prosječna vrijednosti parametra $Q_{2,rnd}$ po lancu; Bin_{uk} – zbroj vrijednosti binomnog koeficijenta pojedinih lanaca u RS; $S_{0,uk}$ – zbroj vrijednosti entropijskog koeficijenta segmentnog nasumičnog modela (bez razmaka između segmenata) pojedinih lanaca za cijeli RS; TM_{sr} – prosječni broj TM segmenata po lancu; $S_{0,sr}$ – srednja vrijednost entropijskog koeficijenta segmentnog nasumičnog modela (kratica 'snm'), bez razmaka između segmenata u RS; $Q_{2,rnd}$ – vrijednost koeficijenta $Q_{2,rnd}$ za cijeli skup; Algoritam 1n (algoritam Hobohm 1 u kojemu je dodan samo nasumični izbor lanca); Algoritam 2n (algoritam Hobohm 2 u kojemu je dodan samo nasumični izbor lanca)

Stoga, na tom skupu u koji su (već) izabrani lanci niske međusobne identičnosti, nisu se mogle pokazati neke značajnije prednosti pojedinih algoritama. Svi su algoritmi dali vrlo slične rezultate (tablica 19) i to po svim analiziranim parametrima. Ipak, i u ovom slučaju, Algoritam 3 postigao je najbolje rezultate po svim kriterijima koji se izdvajaju kao najvažniji, tj. ukupni i prosječni (a) broj TM segmenata (najveći), (b) entropijski koeficijenti (najveći), i (c) nasumična točnost modela s dva stanja $Q_{2,rnd}$ (najniži). Isti rezultati dobiveni su i Algoritmima 2, 3 i 1n.

Za taj je skup (ljubaznošću M. Bernhofera) provedena analiza zalihosti pomoću druge inačice algoritma UniqueProt koja je u međuvremenu objavljena zajedno s izborom novoga reprezentativnog skupa proteinskih lanaca [43] (u tablici 19 označena kao UniqueProt-2). Ukupni broj lanaca (161), TM segmenata (506) i aminokiselina (32294) dobiven s ovom inačicom algoritma osjetno je manji od odgovarajućih vrijednosti u početnom skupu M190 (tablica 2) koji su (redom 190, 569 i 50179). Razliku je uzrokovala različiti tretman dijelova nekih proteinskih lanaca u analizi identičnosti (i dobivanju matrice identičnosti koja je ulaz algoritma UniqueProt [44]) za koji u strukturi nisu navedene koordinate. U novijoj inačici [43], u analizama međusobnih identičnosti takvi dijelovi lanca nisu razmatrani pa je, stoga, možemo reći da je identičnost analizirana strožijim postupkom nego u prvome radu [42].

Možemo reći da su slični rezultati koje su postigli svi razmatrani algoritmi za ovaj skup zapravo ponajprije potvrda ispravnosti njihova rada. Za detaljnije informacije o njihovoj kvaliteti i o razlikama među njima bit će potrebno provesti usporedbu rezultata na zahtjevnijim početnim skupovima proteinskih lanaca.

3.2.2. Rezultati i usporedba algoritama na skupovima S481 i S392

Usporedbu algoritama 1, 2 i 3 razvijenih u disertaciji s originalnim algoritmom nazvanim Hobohm 2 [24] nije bilo moguće provesti stoga što je inačica algoritma dobivena od autora (profesor Uwe Hobohm, Njemačka) radila iznimno sporo, te je usporedba provedena s inačicom algoritma Hobohm 2 razvijenim od strane "Grupe Sydney" u radu [41]. Oni su razvili algoritam Hobohm 2 prema opisu iz literature [24,25] i primijenili ga u izboru reprezentativnog skupa membranskih proteina polazeći od baza OPM i PDB u 2013. godini. Na poslužitelju koji je razvijen uz rad dostupni su njihovi pročišćeni početni skupovi S481 kao i njegov podskup S392 sa strukturama rezolucije ispod 3.5Å. Mrežni poslužitelj nudi mogućnost izbora reprezentativnih skupova na različitim razinama identičnosti i sličnosti u rasponu $20\% - 35\%$. Također, posebno

je izdvojen i njihov odabrani reprezentativni skup S101 iz [42] koji je dobiven uz prag maksimalne sličnosti od 30% uporabom njihove inačice algoritma Hobohm 2.
U nastavku su analizirani rezultati dobiveni sa svim algoritmima i to na skupovima S481 i S382, za četiri vrijednosti praga (razine) identičnosti i sličnosti.

Rezultati dobiveni na skupu S481 za različite pragove identičnosti

Rezultati izbora reprezentativnih skupova različitim algoritmima polazeći od skupa S481 za različite pragove identičnosti u granicama od $20, 25, 30$ i 35% dani su u tablici 20. Ta je analiza provedena kako bi se:

a) provjerila uspješnost pojedinih algoritama promatrano prema parametrima za izabrani skup i prema prosječnim vrijednostima po proteinskom lancu,
b) analizirao odnos između rezultata dobivenih različitim algoritmima i
c) rezultati algoritama razvijenih u disertaciji usporedili s rezultatima dobivenim algoritmom Hobohm 2 iz [42], a koji se mogu dobiti i preuzeti s internetskog poslužitelja (označeni kao Skup – web u tablici 20).

Tablica 20. Osobine reprezentativnih skupova dobivenih primjenom algoritama na početni skup S481 za različite pragove identičnosti.[a]

	lanaca	AK_{uk}	TM_{uk}	$Q_{2,rnd,sr}$	Bin_{uk}	$S_{0,uk}$	TM_{sr}	$S_{0,sr}$	$Q_{2,rnd}$
skup S481 – 20i[b]									
Skup - web	164	47335	820	0.592	26996	4133	5.00	25.20	0.52
Algoritam 1	166	48245	846	0.591	27570	4277	5.10	25.76	0.52
Algoritam 2	164	48184	838	0.591	27286	4226	5.11	25.77	0.52
Algoritam 3	**169**	48221	**850**	0.590	27621	**4284**	5.03	25.35	0.52
skup S481 – 25i[b]									
Skup - web	224	62670	1109	0.584	35359	5517	4.95	24.63	0.52
Algoritam 1	225	62600	1122	0.583	35568	5595	4.99	24.87	0.52
Algoritam 2	223	61902	1120	0.584	35387	5576	5.02	25.00	0.51
Algoritam 3	225	62407	1122	0.583	35511	5586	4.99	24.83	0.51
skup S481 – 30i[b]									
Skup - web	252	67187	1205	0.581	38413	5971	4.78	23.70	0.52
Algoritam 1	254	69292	1219	0.586	38806	6053	4.80	23.83	0.52
Algoritam 2	254	68849	1219	0.585	38643	6037	4.80	23.77	0.52
Algoritam 3	254	69009	1219	0.586	38701	6041	4.80	23.79	0.52
skup S481 – 35i[b]									
Skup - web	276	74170	1309	0.584	41980	6494	4.74	23.53	0.52
Algoritam 1	276	75430	1319	0.586	42195	6556	4.78	23.75	0.52
Algoritam 2	276	75308	1319	0.586	42103	6544	4.78	23.71	0.52
Algoritam 3	276	74674	1319	0.585	41995	6539	4.78	23.69	0.52

[a] oznake kratica dane su ispod tablice 19
[b] 20i = razina/prag 20% identičnosti i analogno za pragove 25, 30 i 35%

Vidi se da je broj izabranih lanaca u reprezentativnim skupovima membranskih proteina alfa vrste sličan za pojedine pragove identičnosti. Slično vrijedi i za druge parametre kvalitete skupa poput broja TM segmenata i broja aminokiselina u skupu, te ukupnog i prosječnog entropijskog koeficijenta segmentnog i binomnog nasumičnog modela. Ipak, uočava se da su

ukupno gledajući nešto malo bolji rezultati dobiveni algoritmima 1, 2, i 3 razvijenim u disertaciji.

Jedina osjetnija razlika je za prag identičnosti od 20% pri kojoj je Algoritam 3 izdvojio 3.05% više lanaca u reprezentativni skup a ti lanci ukupno imaju 3.66% više TM segmenata. Ukupna složenost (kompleksnost) proteinskih struktura reprezentativnog skupa (iskazana ukupnom vrijednošću entropijskog koeficijenta segmentnog nasumičnog modela) izdvojenog Algoritmom 3 veća je za 3.65% od ukupne složenosti dobivene algoritmom Hobohm 2 [42] (Skup - web iz tablice 20). Promatrano po identičnosti, upravo razina identičnosti od 20% najčešće se uzima kao prag pri izboru reprezentativnih skupova, pa je taj rezultat (gledano po identičnosti) i najvažniji.

Rezultati dobiveni na skupu S481 za različite pragove sličnosti

U tablici 21 dan je zbirni prikaz osobina reprezentativnih skupova membranskih proteina alfa vrste izabranih razvijenim algoritmima i algoritmom Hobohm 2 iz [42] za četiri praga sličnosti.

Tablica 21. Osobine reprezentativnih skupova dobivenih primjenom algoritama na početni skup S481 za različite pragove sličnosti.[a]

	lanaca	AK_{uk}	TM_{uk}	$Q_{2,rnd,sr}$	Bin_{uk}	$S_{0,uk}$	TM_{sr}	$S_{0,sr}$	$Q_{2,rnd}$
skup S481 – 20s[b]									
Skup - web	32	9644	129	0.623	5190	678	4.03	21.20	0.57
Algoritam 1	35	10837	135	0.637	5558	721	3.86	20.59	0.59
Algoritam 2	35	10792	141	0.620	5646	745	4.03	21.28	0.58
Algoritam 3	36	11097	158	0.621	5972	832	4.39	23.11	0.55
skup S481 – 25s[b]									
Skup - web	63	19882	255	0.631	10210	1349	4.05	21.42	0.58
Algoritam 1	69	21014	276	0.639	10705	1443	4.00	20.91	0.58
Algoritam 2	64	19759	293	0.607	10886	1536	4.58	24.00	0.55
Algoritam 3	66	21124	297	0.618	11246	1562	4.50	23.67	0.56
skup S481 – 30s[b]									
Skup - web	121	37162	580	0.607	20503	2979	4.79	24.62	0.54
Algoritam 1	125	37717	598	0.602	20979	3080	4.78	24.64	0.54
Algoritam 2	125	37560	604	0.601	20928	3093	4.83	24.74	0.53
Algoritam 3	**127**	38134	**621**	0.602	21330	**3182**	4.89	25.05	0.53
skup S481 – 35s[b]									
Skup - web	178	50237	903	0.584	29071	4517	5.07	25.37	0.51
Algoritam 1	178	50839	908	0.588	29249	4561	5.10	25.62	0.51
Algoritam 2	177	49918	908	0.586	28876	4530	5.13	25.60	0.51
Algoritam 3	177	50001	904	0.585	28935	4520	5.11	25.54	0.51

[a] oznake kratica dane su ispod tablice 19
[b] 20s = razina/prag 20% sličnosti i analogno za pragove 25, 30 i 35%

Ukupno gledajući, Algoritmi 1, 2 i 3 izabrali su kvalitetnije reprezentativne skupove po svim parametrima, osim u slučaju praga sličnosti od 35% za koji su rezultati svih algoritama vrlo slični. Međutim, važniji su rezultati za pragove sličnosti ≤ 30%, jer ispod te sličnosti u primarnim strukturama proteinski lanci ne pokazuju sličnosti u strukturama. Točnije, upravo skupovi proteina međusobne sličnosti ispod 30% su pravi reprezentativni skupovi.

Na standardnoj razini sličnosti od 30% Algoritam 3 dao je rezultate koji su po broju reprezentativnih lanaca veći za ~ 5% u odnosu na rezlutate dobivene algoritmom Hobohm 2 [42], dok je promatrano prema broju transmembanskih segmenata taj rezultat još i bolji (7.07%). Slični se rezultati dobivaju i pri usporedbi ukupnih i prosječnih entropijskih koeficijenta segmentnog nasumičnog modela koji su po ukupnom iznosu za skup veći 6.81%. Kao i na standardnoj razini od 20% identičnosti tako i na standardnoj razini 30% sličnosti pokazuje se da Algoritam 3 izabire veće reprezentativne skupove membranskih proteina složenijih struktura. Dodatno, ako se promotre rezultati izbora skupa za sljedeći prag sličnosti od 25%, rezultati koje postiže Algoritam 3 u usporedbi s algoritmom Hobohm 2 bolji su 16% i prema ukupnom broju TM segmenata u izabranim proteinskim lancima, i složenosti struktura prema ukupnim entropijskim koeficijentom segmentnom nasumičnoga modela.

Rezultati dobiveni na skupu S392 za različite pragove identičnosti

Skup S392 podskup je početnog skupa S481 i za očekivati je da rezultati izbora reprezentativnih skupova različitim algoritmima budu slični onima za skup S481.

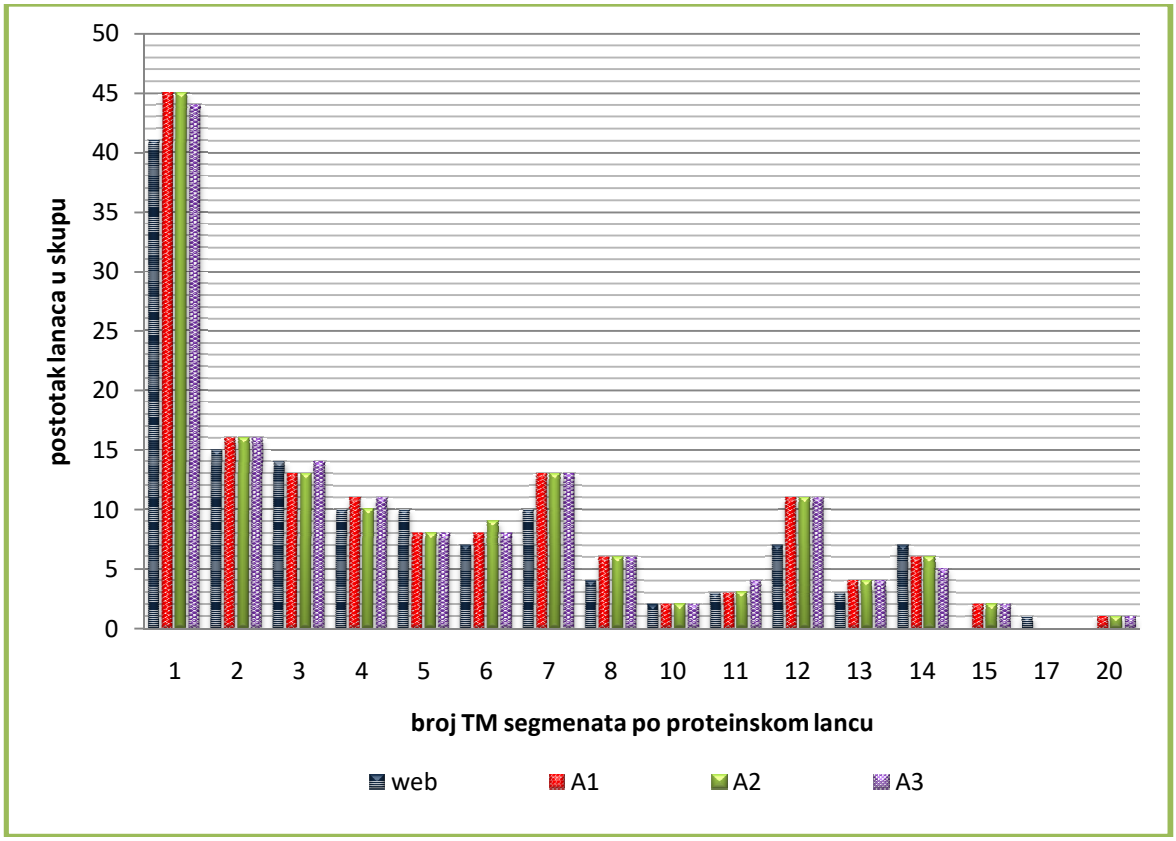
Tablica 22. Osobine reprezentativnih skupova dobivenih primjenom algoritama na početni skup S392 za različite pragove identičnosti.[a]

	lanaca	AK_{uk}	TM_{uk}	$Q_{2,rnd,sr}$	Bin_{uk}	$S_{0,uk}$	TM_{sr}	$S_{0,sr}$	$Q_{2,rnd}$
skup S392 – 20i[b]									
Skup - web	134	36200	638	0.592	20833	3182	4.76	23.75	0.51
Algoritam 1	149	42852	758	0.592	24403	**3814**	5.09	25.60	0.54
Algoritam 2	**149**	41267	**760**	0.585	24091	3800	5.10	25.50	0.51
Algoritam 3	149	42868	757	0.590	24565	3806	5.08	25.54	0.52
skup S392 – 25i[b]									
Skup - web	183	50447	917	0.586	28685	4528	5.01	24.74	0.51
Algoritam 1	200	54552	1001	0.582	31376	4952	5.01	24.76	0.53
Algoritam 2	198	53947	999	0.582	31229	4949	5.05	25.00	0.51
Algoritam 3	200	54597	1001	0.583	31429	4962	5.01	24.81	0.51
skup S392 – 30i[b]									
Skup - web	208	53960	995	0.581	31175	4897	4.78	23.54	0.51
Algoritam 1	224	60596	1096	0.584	34368	5427	4.89	24.23	0.53
Algoritam 2	224	60608	1096	0.583	34419	5423	4.89	24.21	0.51
Algoritam 3	224	60443	1096	0.584	34340	5412	4.89	24.16	0.51
skup S392 – 35i[b]									
Skup - web	227	58974	1075	0.584	33701	5275	4.74	23.24	0.51
Algoritam 1	242	64815	1166	0.585	36639	5758	4.82	23.79	0.53
Algoritam 2	242	64705	1166	0.584	36593	5754	4.82	23.78	0.51
Algoritam 3	242	64135	1166	0.583	36548	5741	4.82	23.72	0.51

[a] oznake kratica dane su ispod tablice 19
[b] 20i = razina/prag 20% identičnosti i analogno za pragove 25, 30 i 35%

Zapravo, ovakva vrsta provjere pokazat će osjetljivost algoritama na prethodnu pripremu i pročišćavanje skupa proteinskih lanaca pohranjenih u proteinskim bazama podataka. Inače, to je obvezan korak u ovakvim analizama, i taj korak može utjecati na rad algoritma za izbor reprezentativnih skupova i na samu kvalitetu tih skupova.

Rezultati dobiveni na skupu S392 za četiri praga identičnosti ponovno pokazuju da su Algoritmi 1, 2 i 3 izabrali kvalitetnije i složenije reprezentativne skupove u odnosu na rezultate dobivene algoritmom Hobohm 2 iz ref. [42]. Na standardnoj razini (pragu) identičnosti između proteinskih lanaca od 20%, brojevi lanaca u izabranim skupovima dobiveni Algoritmima 1, 2 i 3 viši su za 11.19%. Dodatno, u reprezentativnom skupu izabranom Algoritmom 2 broj TM segmenata viši je za 19.12%, dok je u skupovima izabranim sa svakim od tri algoritma razvijenim u disertaciji broj TM segmenata u skupu iznad 18.65%. Ukupna složenost struktura u skupovima izabranim tim algoritmima (prema entropijskom koeficijentu segmentnog nasumičnog modela) viša je od 19% u odnosu na rezultat dobiven algoritmom Hobohm 2. Iako se radi o podskupu skupa S481, razlike su se povećale u korist algoritama razvijenih u disertaciji.



Slika 27. Raspodjela broja lanaca (u postotcima) po podskupovima lanaca istog broja TM segmenata u početnom skupu S392 i u reprezentativnim skupovima izabranim Algoritmima 1, 2, 3 i Hobohm 2 (uz prag identičnosti 20%).

Na slici 27 prikazane su raspodjele postotnih udjela lanaca u reprezentativnim skupovima iz tablice 22 u ovisnosti o broju TM segmenata. Svi skupovi izabrani su algoritmom Hobohm 2 i Algoritmima 1, 2 i 3 iz početnog skupa S392 uz prag identičnosti 20%. Uočava se veći postotni udio lanaca s 1, 7 i 12 TM segmenata u reprezentativnim skupovima izabranim Algoritmima 1, 2 i 3 u odnosu na skup izabran algoritmom Hobohm 2.

Rezultati dobiveni na skupu S392 za različite pragove sličnosti

Rezultati dobiveni na skupu S392 za četiri praga sličnosti pokazuju da su Algoritmi 1, 2 i 3 izabrali kvalitetnije i složenije reprezentativne skupove u odnosu na rezultate dobivene algoritmom Hobohm 2 iz ref. [42]. U tablici 22 dani su rezultati za četiri praga sličnosti, a najvažniji je među njima standardni prag od 30% (najčešće korišten za sličnost). Za prag 30% podebljani su najbolji rezultati prema parametima složenosti i kvalitete skupa.

Tablica 23. Osobine reprezentativnih skupova dobivenih primjenom algoritama na početni skup S392 za različite pragove sličnosti.[a]

	lanaca	AK_{uk}	TM_{uk}	$Q_{2,rnd,sr}$	Bin_{uk}	$S_{0,uk}$	TM_{sr}	$S_{0,sr}$	$Q_{2,rnd}$
skup S392 – 20s[b]									
Skup - web	23	6195	91	0.633	3389	467	3.96	20.29	0.55
Algoritam 1	35	10152	134	0.638	5209	706	3.83	20.17	0.57
Algoritam 2	32	10278	137	0.633	5263	716	4.28	22.38	0.56
Algoritam 3	33	10085	149	0.626	5402	784	4.52	23.74	0.56
skup S392 – 25s[b]									
Skup - web	51	13998	185	0.630	7272	959	3.63	18.8	0.57
Algoritam 1	63	18777	268	0.621	10108	1390	4.25	22.06	0.55
Algoritam 2	63	18569	255	0.619	9868	1321	4.05	20.97	0.56
Algoritam 3	66	19316	299	0.612	10659	1547	4.53	23.44	0.54
skup S392 – 30s[b]									
Skup - web	101	30144	483	0.603	16899	2462	4.78	24.38	0.53
Algoritam 1	**118**	33467	555	0.600	18866	2814	4.7	23.84	0.52
Algoritam 2	115	32684	**564**	0.589	18943	**2855**	4.9	24.82	0.52
Algoritam 3	117	33262	560	0.593	19108	2846	4.79	24.33	0.52
skup S392 – 35s[b]									
Skup - web	148	39642	736	0.582	23303	3650	4.97	24.66	0.51
Algoritam 1	160	44339	807	0.586	25882	4040	5.04	25.25	0.51
Algoritam 2	160	43668	809	0.582	25564	4025	5.06	25.16	0.51
Algoritam 3	160	43694	809	0.583	25649	4024	5.06	25.15	0.51

[a] oznake kratica dane su ispod tablice 19
[b] 20s = razina/prag 20% sličnosti i analogno za pragove 25, 30 i 35%

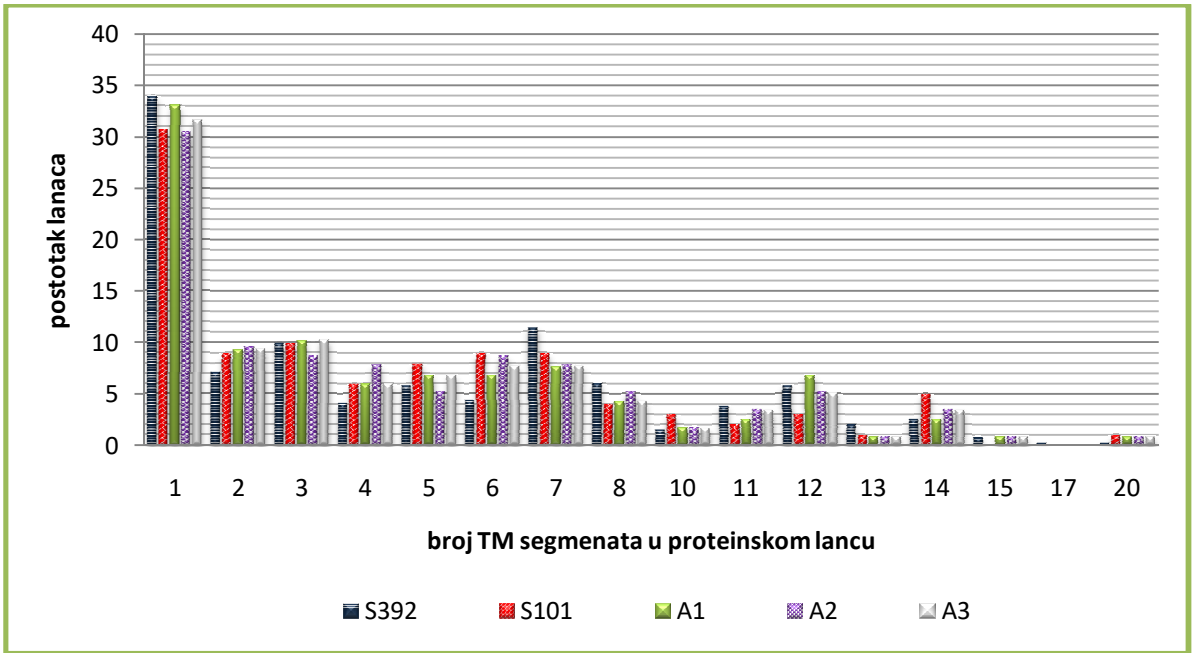
Najveća je razlika između dobivenih skupova za prag sličnosti 20%. Najveća je razlika u broju lanaca između skupova dobivenih algoritmom Hobohm 2 i Algoritmom 1, i iznosi 52.17%. I druga dva algoritma daju preko 39% više lanaca u odnosu na algoritam Hobohm 2 [42]. Nadalje, broj TM segmenata veći je za preko 47.25% sa sva tri algoritma, a najbolji je rezultat dobiven Algoritmom 3 kod kojeg je broj TM segmenata veći 63.74% u usporedbi s algoritmom Hobohm 2. Još je veća razlika u ukupnoj složenosti struktura prema entropijskom koeficijentu segmentnog nasumičnog modela rezentativnih skupova, gdje su strukture izabrane Algoritmom 3 ukupno gledano složenije 67.88%.

Na standardnoj razini sličnosti od 30% na kojoj je "Grupa Sydney" algoritmom Hobohm 2 dobila reprezentativni skup S101 sa 101 lancem, Algoritmi 1, 2 i 3 izabrali su od 115 do 118 lanaca. Taj je rezultat bolji 13.86-16.83% u odnosu na skup S101 iz [41]. Broj TM segmenata izabran Algoritmima 1, 2 i 3 veći je 14.27-15.94%, dok je ukupna složenost struktura prema entropijskom koeficijentu segmentnog nasumičnog modela veća 14.27-15.94%.

Iz podataka u tablici 23 primjećuje se da se razlike u reprezentativnim skupovima koji su izabrani Algoritmima 1, 2 i 3 u odnosu na standardni algoritam Hobohm 2 smanjuju s porastom praga sličnosti. Za prag sličnosti 30%, razlike između parametara standardnog skupa S101 izabranog u [41] i skupova dobivenih Algoritmima 1, 2 i 3 iznose 16.83% u broju lanaca, 16.77% u ukupnom broju TM segmenata, te 15.94% po ukupnoj složenosti struktura prema $S_{0,uk}$ (entropijski koeficijent segmentnog nasumičnog modela). Najmanja razlika u broju izabranih lanaca je za prag sličnosti 35% i iznosi 8.11%. Ako se ima na umu podatak da broj novih

jedinstvenih proteinskih lanaca u bazi PDB (koji imaju nisku sličnost s postojećim proteinskim lancima membranskih proteina poznate strukture) raste jako sporo (osjetno ispod 8 ili 16% gledano i kroz pet godina), jasno je kako novi razvijeni algoritmi značajno unapređuju izbor reprezentativnih skupova.

Na slici 29 prikazane su raspodjele postotnih udjela lanaca u reprezentativnim skupovima iz tablice 23 u ovisnosti o broju TM segmenata. Svi skupovi izabrani su algoritmom Hobohm 2 i Algoritmima 1, 2 i 3 iz početnog skupa S392 uz prag sličnosti 30%.



Slika 28. Raspodjela broja lanaca (u postotcima) po podskupovima lanaca istog broja TM segmenata u početnom skupu S392 i u reprezentativnim skupovima izabranim Algoritmima 1, 2, 3 i Hobohm 2 (uz prag sličnosti 30%).

Uočava se pad postotnog udjela lanaca s 1, 7 i 8 i porast udjela lanaca s 2, 4 i 6 TM segmenata u svim reprezentativnim skupovima u odnosu na početni skup S392. Nadalje, postotak proteinskih lanaca u reprezentativnom skupu S101 u odnosu na skupove izabrane Algoritmima 1, 2 i 3 veći je za lance s 5, 6, 7, 10 i 14 TM segmenata a manji samo za lance s 11, 12 i 15 TM segmenata. S obzirom da je prikazan postotni udio koji je normiran s obzirom na veličinu skupa, iz ovoga ne možemo izvlačiti zaključke o ukupnim brojevima TM segmenata u reprezentativnim skupovima. Uglavnom, uočava se kako se raspodjele lanaca u ovisnosti o broju TM segmenata u njima nisu značajno mijenjale u analizama za prag sličnosti 30% ni u odnosu na početni skup S392, a niti između skupova izabranim raznim algoritmima.

3.2.3. Rezultati dobiveni na skupu M1087

Primjenom Algoritama 1, 2, 3 razvijenih u disertaciji i algoritma UniqueProt [43] na veći početni skup M1087 dobiveni su rezultati prikazani u tablici 24. Vidi se da je Algoritam 3 izabrao reprezentativni skup najveće ukupne složenosti (prema entropijskom koeficijentu segmentnog nasumičnog modela $S_{0,uk}$) koja je za 1% veća od dobivene ukupne složenosti dobivene algoritmom UniqueProt [43]. Dodatno, prosječni broj TM segmenata u skupu dobivenom Algoritmom 3 veći je za 3.69%, unatoč 2.41% manjem broju proteinskih lanaca u odnosu na skup izabran algoritmom UniqueProt. Prosječni broj TM segmenata po proteinskom lancu u

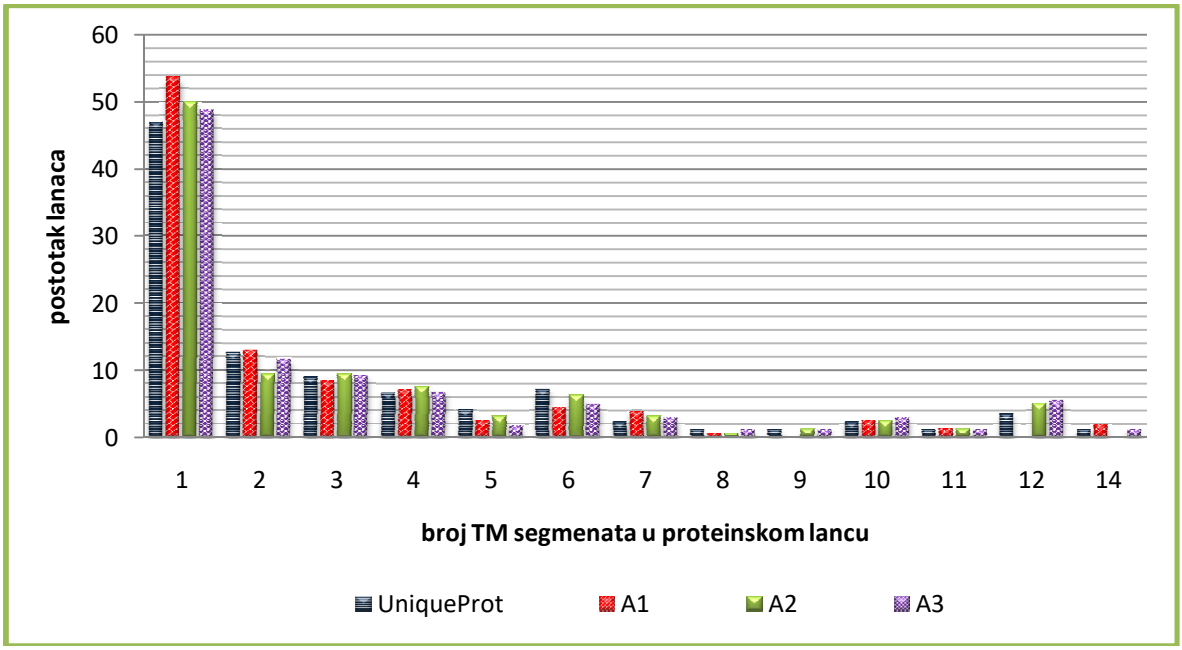
početnom skupu M1087 je 4.70. Nakon izbacivanja potpuno identičnih lanaca, taj prosjek iznosi 4.91.

Tablica 24. Osobine reprezentativnih skupova dobivenih primjenom algoritama na početni skup M1087 uz prag identičnosti 20%.[a]

	lanaca	AK_{uk}	TM_{uk}	$Q_{2,rnd,sr}$	Bin_{uk}	$S_{0,uk}$	TM_{sr}	$S_{0,sr}$	$Q_{2,rnd}$
				skup M1087 – 20i[b]					
UniqueProt	**166**	33301	540	0.60	17655	2529	3.25	15.24	0.55
Algoritam 1	153	28217	421	0.61	14236	1947	2.75	12.73	0.57
Algoritam 2	158	30995	506	0.60	16242	2342	3.20	14.82	0.55
Algoritam 3	162	33764	**546**	0.60	17371	**2555**	3.37	15.77	0.55

[a] oznake kratica dane su ispod tablice 19
[b] 20i = razina/prag 20% identičnosti

Algoritmom UniqueProt [43] izabrano je 166 proteinskih lanaca u reprezentativni skup, što je najbolji rezultat gledano prema broju lanaca. No, najveći broj TM segmenata, kao i najveću složenost imaju skupovi izabrani Algoritmom 3.



Slika 29. Raspodjela broja lanaca (u postotcima) po podskupovima lanaca istog broja TM segmenata u početnom skupu M1087 i u reprezentativnim skupovima izabranim Algoritmima 1, 2 i 3 (uz prag identičnosti 20%).

Na slici 29 prikazane su raspodjele postotnih udjela lanaca u reprezentativnim skupovima iz tablice 24 u ovisnosti o broju TM segmenata.

Svi skupovi izabrani su algoritmom UniqueProt [43] i Algoritmima 1, 2 i 3 iz početnog skupa M1087, uz prag identičnosti 20%. Uočavamo veliku ujednačenost u raspodjeli lanaca u ovisnosti o broju TM segmenata. Jedino značajnije odstupanje vezano je uz mali porast broja lanaca s jednim TM segmentom u reprezentativnim skupovima izabranim Algoritmima 1, 2 i 3 u odnosu na skup izabran algoritmom UniqueProt [43].

3.2.4. Rezultati dobiveni na skupu S148

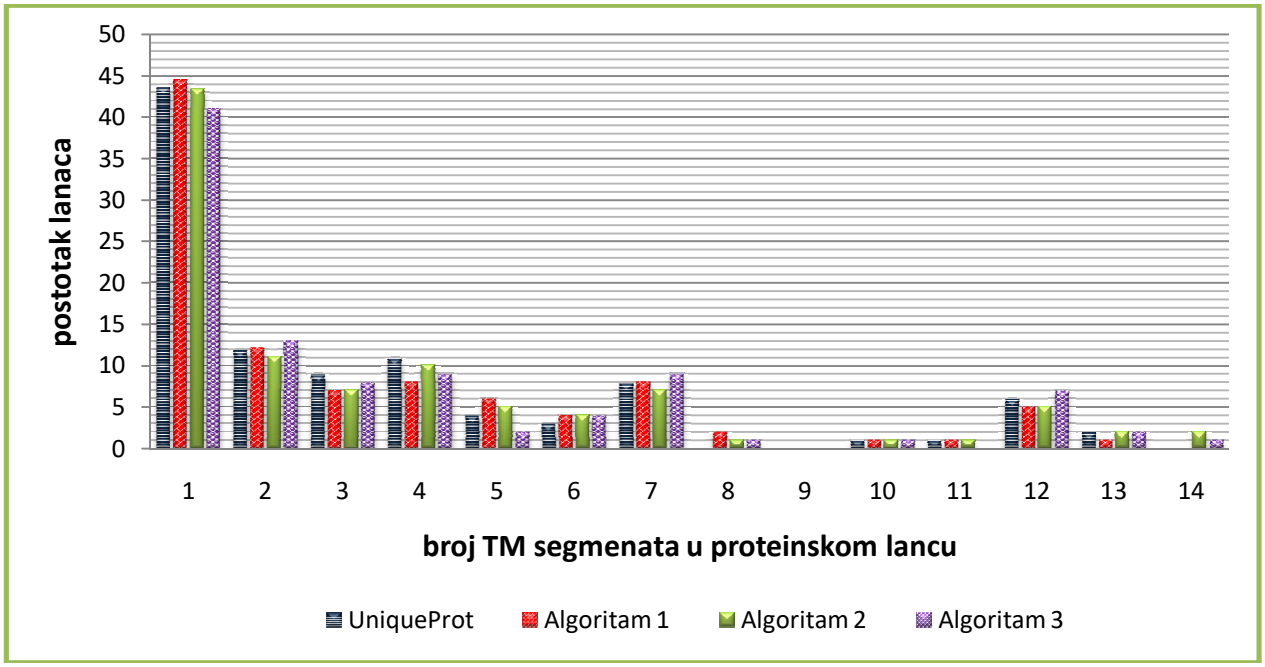
Rezultati izbora reprezentativnih skupova raznim algoritmima prikazani su u tablici 24. Vidi se da je algoritam UniqueProt izabrao 101 reprezentativni lanac od početnih 148, dok je Algoritam 1n (što je zapravo algoritam koji radi slično algoritmu Hobohm 1 u kojem je uključen nasumični izbor lanca pri odabiru) izabrao najviše proteinskih lanaca.

Tablica 25. Parametari kvalitete reprezentativnih skupova izabranih iz skupa S148.[a]

	lanaca	AK_{uk}	TM_{uk}	$Q_{2,rnd,sr}$	Bin_{uk}	$S_{0,uk}$	TM_{sr}	$S_{0,sr}$	$Q_{2,rnd}$
					skup S148 – 20i[b]				
UniqueProt	101	21586	352	0.599	11716	1681	3.49	16.64	0.55
Algoritam 1	99	20664	341	0.600	11178	1622	3.44	16.39	0.55
Algoritam 2	99	21764	367	0.599	11912	1762	3.71	17.8	0.54
Algoritam 3	100	22855	**396**	0.595	12649	**1913**	**3.96**	19.13	0.54
Hobohm 1	100	20346	329	0.602	10863	1550	3.29	15.5	0.55
Hobohm 2	98	20195	345	0.596	11119	1631	3.52	16.64	0.6
Algoritam 1n	102	21420	359	0.597	11781	1703	3.52	16.7	0.54
Algoritam 2n	101	21485	358	0.600	11700	1706	3.54	16.9	0.54

[a] oznake kratica dane su ispod tablice 19
[b] 20i = razina/prag 20% identičnosti

Unatoč tome što je Algoritam 3 odabrao manje lanaca (100) u reprezentativni skup, ukupni broj TM segmenata (tj. 396) znatno je iznad ostalih skupova. To je za 7.9% više od drugog algoritma najboljeg po tom kriteriju, a od broja TM segmenata u skupu izabranom algoritmom UniqueProt (352), taj rezultat bolji je za 12.5%. Nadalje, može se vidjeti u tablici 24, da je ukupna složenost modelnih struktura skupa dobivenog Algoritmom 3 veća za 13.8% od složenosti skupa izabranog algoritmom UniqueProt.



Slika 30. Raspodjela broja lanaca (u postotcima) po podskupovima lanaca istog broja TM segmenata u početnom skupu S148 i u reprezentativnim skupovima izabranim Algoritmima 1, 2 i 3 (uz prag identičnosti 20%).

Na slici 30 prikazan je postotni udio proteinskih lanaca u reprezentativnim skupovima proteina dobivenim primjenom algoritama na skupu S148, za prag identičnosti 20%. Vidimo opet veliku ujednačenost raspodjela, te se uočava da nijedan algoritam nije izabrao proteinske lance s 15 i 20 TM segmenata. Uočava se da Algoritam 3 bira nešto manje proteinskih lanaca s jednim TM segmentom, a nešto više u skupinama lanaca s 2, 7 i 12 TM segmenata. U konačnici, taj izabrani reprezentativni skup uključuje ukupno najsloženije strukture, koje su i najteže za analizu i predviđanje.

3.2.5. Rezultati na skupovima N189 i N263

Skupovi N189 i N263 reprezentativni su skupovi izabrani na temelju matrice identičnosti dobivene programom EMBOSS [49] uz prag 18% i 20% (redno), i prema toj matrici, izabrani lanci u svakom od skupova nisu identični s drugim lancima više od 18% (N189) odnosno 20% (N263). Međutim, s obzirom da se identičnost između proteinskih lanaca prema metodi UniqueProt [44] računa nešto drugačije (što je opisano u 2.3.2), za ove skupove izračunana je matrica identičnosti prema UniqueProt (ljubaznošću M. Bernhofera, jedan od autora rada [43]) i polazeći od te matrice provedeni su izbori reprezentativnih skupova uz prag identičnosti 20%:

a) algoritmom Uniqueprot [44] prema postupku opisanom u [43] (analizu je proveo M. Bernhofer),
b) algoritmima Hobohm 1 i 2 te Algoritmima 1, 2 (po dvije inačice) i 3 razvijenim u disertaciji uz prag identičnosti 20%.

Usporedbe na ova dva skupa posebne su po tome što se provodi unakrsna provjera kvalitete algoritama strogosti jednog i drugog načina računanja identičnosti. Naime, želi se provjeriti koliko je izbor po identičnosti računanoj prema metodi EMBOSS [49] optimalan s obzirom na identičnost računanoj metodom UniProt [44]. Rezultati dobiveni na skupu N189 opisanom u 2.4.5 za prag identičnosti 20% pokazuju da su Algoritmi 2 i 3 izabrali kvalitetnije i složenije reprezentativne skupove u odnosu na algoritam UniqueProt i druge algoritme (tablica 26). Ovi algoritmi izabrali su skupove s najvećim ukupnim (i prosječnim) brojevimaTM segmenata i ukupnim (i prosječnom) složenostima struktura u izabranim skupovima (prema entropijskom koeficijentu segmentnog nasumičnog modela). Nadalje, skupovi izabrani tim algoritmima imaju najnižu ukupnu razinu slučajne točnosti modela s dva stanja (iznos $Q_{2,rnd} = 0.55$). Međutim, zanimljivo je pritom uočiti da je broj lanaca u skupovima izabranim Algoritmom 1 i 2 najmanji (115).

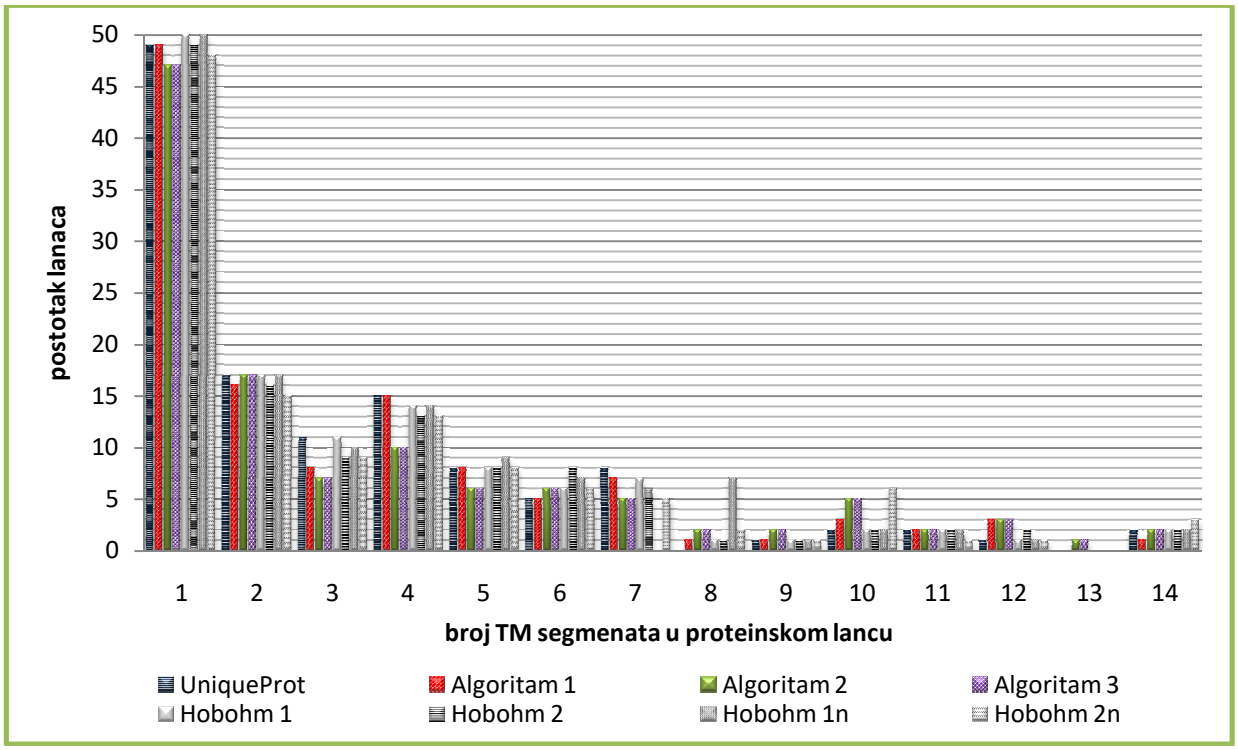
Tablica 26. Osobine reprezentativnih skupova dobivenih primjenom algoritama na početni skup N189 za prag identičnosti 20%.

	lanaca	AK_{uk}	TM_{uk}	$Q_{2,rnd,sr}$	Bin_{uk}	$S_{0,uk}$	TM_{sr}	$S_{0,sr}$	$Q_{2,rnd}$
skup N189 – 20i[b]									
UniqueProt	121	25813	393	0.616	13402	1853	3.25	15.31	0.56
Algoritam 1	119	26172	403	0.615	13687	1907	3.39	16.03	0.56
Algoritam 2	115	26564	**426**	0.613	14137	**2052**	**3.70**	**17.84**	**0.55**
Algoritam 3	115	26564	**426**	0.613	14137	**2052**	**3.70**	**17.84**	**0.55**
Hobohm 1	122	26726	397	0.618	13570	1874	3.25	15.36	0.57
Hobohm 2	119	25293	401	0.613	13430	1903	3.37	15.99	0.56
Algoritam 1n	122	26936	397	0.617	13732	1888	3.25	15.47	0.57
Algoritam 2n	118	24947	418	0.608	13458	1967	3.54	16.67	0.55

[a] oznake kratica dane su ispod tablice 19
[b] 20i = razina/prag 20% identičnosti

Usporedimo li broj lanaca u izabranim skupovima s brojem u početnom skupu (N189), vidi se da je smanjenje u prosjeku 37%, što ukazuje na velike razlike u metodama EMBOSS [49] i UniqueProt [49] u načinu kvantificiranja identičnosti. Ovaj rezultat može se usporediti s onim iz tablice 19 u kojoj su dane osobine izabranih skupova kada je početni skup bio reprezentativni skup M190 sa 190 lanaca izabran metodom UniqueProt [42,44], a prosječno smanjenje broja lanaca iznosi oko 16%. Međutim, i kod tog skupa razvijeni algoritmi dali su bolji rezultat kada se promatraju parametri koji su u vezi sa složenosti strukture.

Na slici 31 prikazana je raspodjela broja lanaca u ovisnosti o broju TM segmenata u reprezentativnim skupovima izabranih iz početnog skupa N189 polazeći od matrice identičnosti računane metodom UniqueProt uz prag identičnosti 20%. Početni skup N189 ima po jedan lanac sa 16 i 17 (nije prikazano na slici 31) TM segmenata, dok niti jedan algoritam u konačnom skupu nije izabrao te lance. Ovaj rezultat vodi na zaključak da odabir lanaca s najvećim brojem TM segmenata ne vodi nužno ka odabiru reprezentativnog skupa najveće složenoti struktura, odnosno s najvećim brojem lanaca i TM segmenata.



Slika 31. Raspodjela broja lanaca (u postotcima) po podskupovima lanaca istog broja TM segmenata u reprezentativnim skupovima izabranim s više algoritama iz skupa N189 (uz prag identičnosti 20%).

U skupu N263 nalaze se 64 jedinstvena proteinska lanca, to znači da nemaju niti jednog susjeda odnosno lanca s kojim su identični više od 20%. Prema tome, izbor konačnog skupa ovisi o preostalih 199 proteinskih lanaca. Konačni rezultati dobiveni primjenom različitih algoritama prikazani su u tablici 27 gdje su podebljane vrijednosti dobivene Algoritmom 3 (dva rješenja). Jedno rješenje daje najveći broj lanaca, a drugo najveći ukupni broj TM segmenata i najveću ukupnu konformacijsku entropiju, odnosno složenost struktura reprezentativnog skupa. Prema tome, vidi se da skup s najvećim brojem proteinskih lanaca nije nužno i najsloženiji skup.

Usporedba s algoritmom UniqueProt [43,44] pokazuje da se u skupu izabranom Algoritmom 3 nalazi 1 lanac više i 112 TM segmenata više (ili 12.42%), dok je ukupna složenost viša za 12.81%. Iz tablice 27 uočava se kako je prema broju TM segmenata Algoritmom 3 dobiven bolji rezultat zbog izbora proteinskih lanaca koji pripadaju skupinama lanaca s 10, 12 i 13 TM segmenata (podebljano za Algoritam 3 u tablici 27). Vidi se da Algoritam 3 izabire u

reprezentativni skup veći broj lanaca iz podskupova lanaca s većim brojem TM segmenata. Takvim lancima Algoritam 3 daje veću važnost (viši prioritet) pri izboru nego lancima koji imaju veću duljinu, a manji broj TM segmenata.

Tablica 27. Broj lanaca s istim brojem TM segmenata u početnom skupu N263 i u reprezentativnim skupovima izabranim s devet algoritama.

Algoritam / TM segmenata po lancu	*N263*	*UP*	*A1*	*A2*	*A3mp*	*A3*	*H1*	*H2*	*H1n*	*H2n*
1	68	59	59	58	58	**57**	61	57	61	59
2	24	19	17	16	18	19	17	18	18	16
3	19	14	12	10	11	11	12	10	11	12
4	23	17	17	15	16	14	14	18	13	19
5	14	12	11	12	11	9	11	11	10	10
6	21	9	10	12	8	6	11	7	10	6
7	20	8	6	8	7	8	5	7	6	8
8	9	1	2	1	4	3	3	3	3	3
9	5	1	3	1	3	2	3	1	4	2
10	11	2	1	5	5	**5**	3	1	5	1
11	6	1	2	0	1	1	0	2	0	2
12	21	2	3	3	4	**7**	4	3	3	3
13	6	0	0	1	2	**3**	0	0	1	0
14	11	1	1	2	2	2	2	1	2	2
15	2	0	0	0	0	0	0	0	0	0
16	1	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0
18	1	0	0	0	0	0	0	0	0	0
19	1	0	0	0	0	0	0	0	0	0
Ukupno lanaca	263	146	144	144	**150**	**147**	146	139	147	143
Ukupno TM	1461	463	479	512	**565**	**575**	500	456	521	483
Ukupno S_0	7375	2121	2215	2384	**2647**	**2717**	2306	2116	2415	2233

UP – UniqueProt; A1, A2 i A3 – Algoritmi 1, 2 i 3 (opisani u 3.1.6. i 3.1.7.); A3mp – Algoritam 3 (inačica A3 koja daje najveći broj lanaca); A3 – Algoritam 3; H1, H2 (opisani u 2.3.1.), H1n i H2n – algoritmi Hobohm 1, 2, 1n i 2n ('n' je oznaka za inačicu algoritma s nasumičnim odabirom) .

Sličan zaključak daje i usporedba reprezentativnih skupova dobivenih primjenom algoritama H1, H2, H1n i H2n. Algoritmom Hobohm 1n (H1n) izabran je približno isti broj proteinskih lanaca kao i s ostale tri inačice srodnih algoritama, ali s većim ukupnim brojem TM segmenata i većom složenošću struktura (tj. većim entropijskim koeficijentom).

3.3. Reprezentativni skupovi transmembranskih proteina alfa tipa izabrani u disertaciji

Ukoliko se prema definiranim kriteriijima nastoji izdvojiti neki podskup membranskih proteina, poput reprezentatvnog (pod)skupa integralnih membranskih proteina alfa vrste niske međusobne sličnosti, u pravilu se polazi od specijaliziranih baza membranskih proteina poput OPM [29] i PDBTM [55,56]. Te baze dodatno definiraju posebna svojstva membranskih proteina, i optimiraju njihovo smještanje u membranu, dajući dodatne informacije o dijelu lanca koji tvore

TM segmente i o topologiji membranskog proteina. Baza OPM sve je više u uporabi od njenog pojavljivanja 2006. godine, te se odlučilo koristiti informacije i podatke iz te baze u disertaciji. Nakon provođenja optimizacije algoritama na radnim skupovima i skupovima iz literature, potrebno je na kraju izabrati reprezentativne skupove polazeći od najnovijih inačica baza PDB [20] i OPM [29], i predložiti ih za korištenje u razvoju novih metoda za predviđanje strukture membranskih proteina.

Taj dio istraživanja, koliko god se čini dobro definiran, nije jednostavan niti jednoznačan. Zbog prisutnih nesavršenosti i neusklađenosti zapisa struktura, dolazi do pogrešaka u izabranim skupovima struktura, i na pronalaženje i ispravljanje tih pogrešaka potrebno je potrošiti puno vremena. S ciljem što učinkovitijeg ispravljanja takvih pogrešaka, u narednom periodu planirano je provesti i dodatno uparivanje i provjeravanje informacija između baza OPM [29] i PDBTM [55,56]. To se prije svega odnosi na razlučivanje nejasnoća i nepreciznosti (npr. u položajima TM segmenata u lancu) u pojedinim strukturama membranskih proteina. Dodatno, uparili bi se i zapisi primarne strukture iz baze PDB [20] s odgovarajućim zapisima u bazi UniProt [32].

3.3.1. Početni skup lanaca N1212 i njegov podskup N907

Kako bi se izabrali reprezentativni skupovi integralnih membranskih proteina alfa vrste, iz baze podataka OPM izdvojeni su svi proteinski lanci iz skupa membranskih proteina (podskupine bitopic i polytopic) s TM segmentima u sekundarnoj strukturi alfa uzvojnice [29]. Nakon toga odabrani su lanci koji su predstavnici podskupina identičnih lanaca iz baze OPM, i njima su pridruženi aminokiselinski slijedovi iz baze podataka PDB [20]. Naime, mnogo membranskih proteina sastoji se od više identičnih lanaca iste sekundarne strukture, pa se među takvim proteinima izabire samo jedan lanac kao predstavnik.

3.3.2. Opis izbora skupova

U izdvajanju početnog skupa N1212 iz baza OPM [29] i PDB [20] u prvi skup izabrano je 1424 proteina od kojih je njih 379 čije podjedinice jednom prolaze kroz membranu (bitopic) i 1045 onih čije podjedinice prolaze kroz membranu dva ili više puta (polytopic). Zatim je izdvojen popis (lista) s 4348 proteinskih lanaca kojima su određeni položaji TM segmenata. Uparivanjem proteinskih kodova i oznaka lanaca u ta dva zapisa, skupu aminokiselinskih slijedova proteinskih lanaca pridružena je informacija o položaju TM segmenata u slijedu. Pritom nisu uzimani u obzir poli-alaninski lanci, tj. oni koji su sadržavali u slijedu samo aminokiselinu alanin. Zatim je skupu pridružena informacija o rezoluciji s kojom je riješena struktura pojedinog proteina, i skup se slaže (uređuje, sortira) prvo po nazivu, pa po rezoluciji (pri vrhu su lanci najniže rezolucije), te se izuzima samo po jedan predstavnik iz svakog podskupa identičnih lanaca. Tako se došlo do 1815 lanaca iz 1046 struktura membranskih proteina.

Potom se svakom lancu pridružila primarna struktura iz baze PDB [20]. Pri tom nisu uzimani u obzir proteinski slijedovi koji su:

a) sadržavali 'X' kao oznaku za neku nepoznatu aminokiselinu,
b) imali strukture koju su u ključnoj riječi (engl. *Structure Title*) vezano za naziv strukture sadržavale '*Model*', '*Models*' ili '*MODEL*', kao i '*chimera*' ili '*chimeric*',
c) sadržavali manje od 15 aminokiselina.

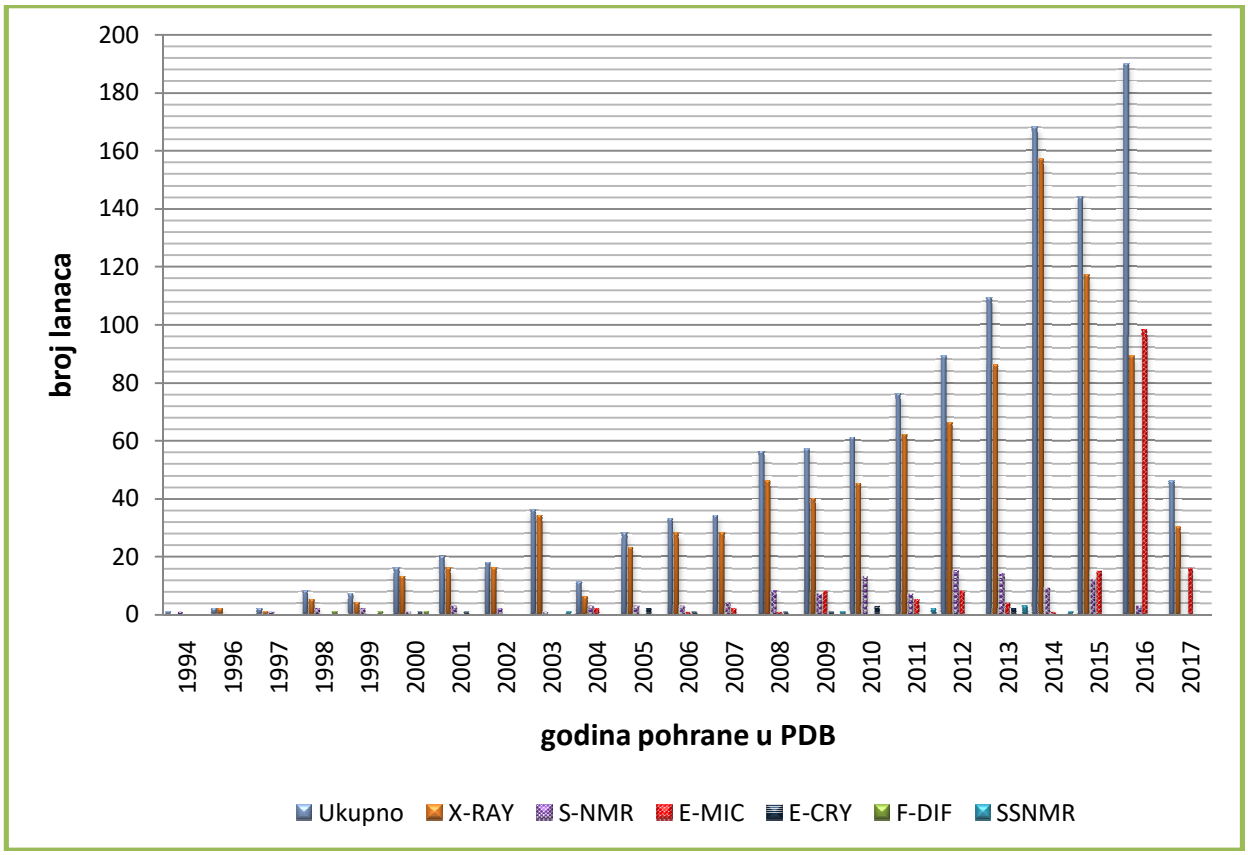
Nakon toga preostalo je 1212 proteinskih slijedova koji čine početni skup nazvan kraće N1212.

Potom je iz tog skupa izdvojen podskup s 907 lanaca nazvan N907 koji sadrži samo lance čija je struktura određena eksperimentalnim metodama rendgenske difrakcije s rezolucijom do 3.5Å i nuklearne magnetske rezonancije u otopini. Radi usporedbe, ovaj je podskup izdvojen u analogiji s podskupom lanaca S392 [41] opisanim u dijelu 2.4.2.

3.3.3. Analize odabranih lanaca i njihovih struktura

S obzirom da je dobiveno znatno više lanaca u početnom skupu u odnosu na početne skupove drugih autora iz 2013. [41] i 2015. [42] godine, ali i ~ 130 lanaca više nego u početnom skupu (skup M1087 opisan u 2.4.3) u novom radu iz 2016. godine [43], analizirani su lanci u skupovima prema vremenu njihovog pohranjivanja u bazu PDB (i OPM). Rezultati su prikazani na slici 32 za skup N1212 i na slici 33 za skup N907, a razvrstani su po: (a) godini pohrane u bazu PDB i (b) eksperimentalnoj metodi kojom je određena struktura lanca.

.



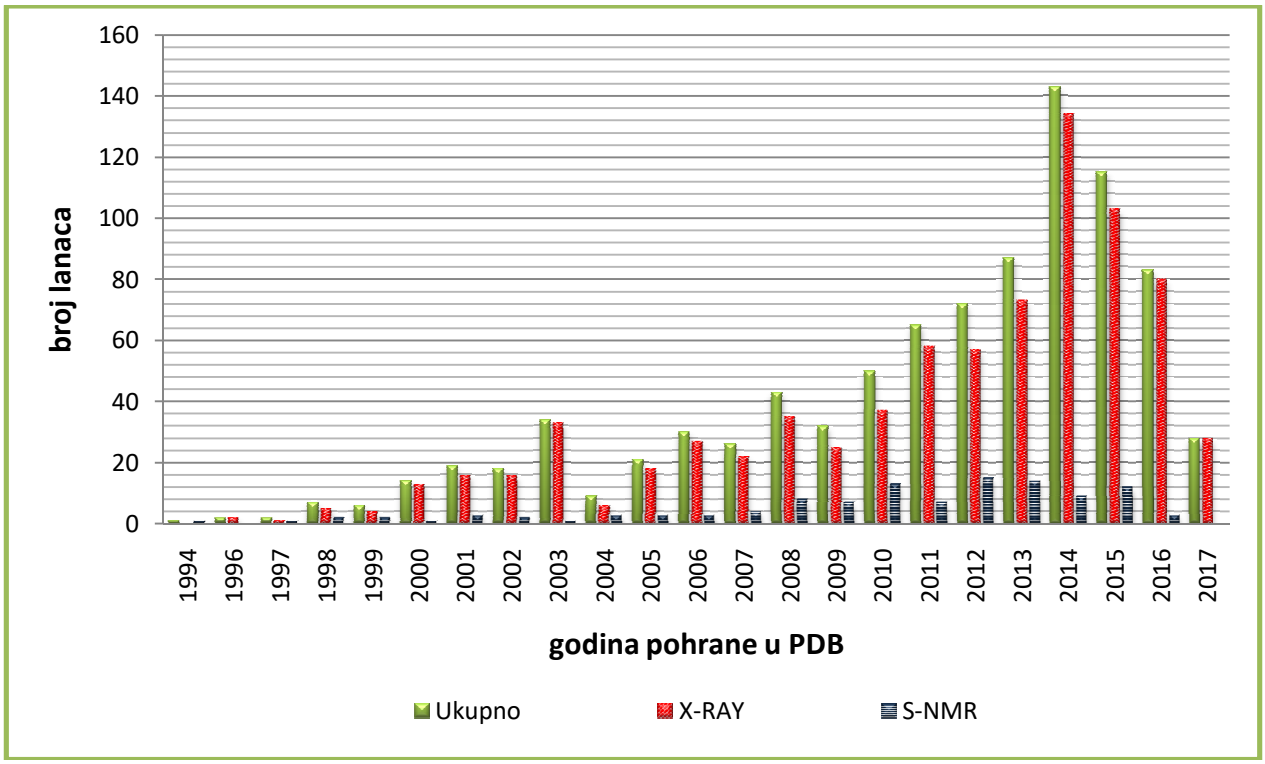
Slika 32. Broj lanaca po godinama pohrane u bazu PDB i po eksperimentalnim metodama kojima je određena struktura u skupu N1212.

Vidi se da je 54.21% lanaca iz skupa N1212 pohranjeno u PDB nakon početka 2013. godine, a njih 45.21% nakon početka 2014. Iz toga slijedi da oko polovice lanaca iz skupa N1212 nije moglo biti uključeno u prethodno navedenim skupovima drugih autora iz disertacije [41,42]. Analizom broja lanaca iz skupa N1212 koji su zastupljeni u ranije objavljenim skupovima S481 i M1087, dobiveni su sljedeći rezultati:

a) skup N1212 ima 315 lanaca iz skupa S481 (65.49%),
b) skup N1212 ima 418 lanaca iz skupa M1087.

Dodatno, uočava se veliki porast broj lanaca kojima je struktura određena elektronskom krio-mikroskopijom u 2016. godini, kada je ovom metodom određena struktura za čak 100 lanaca, što je veliki porast u odnosu na 2015. godinu kada je bilo samo 15 takvih struktura. Dodatno, u 2016. godini broj struktura određenih metodom rendgenske difrakcije (engl. *X-ray*) bio je manji od broja struktura određenih elektronskom krio-mikroskopijom. Taj se trend nastavlja i u 2017. godini. Zanimljivo je spomenuti kako je 2017. godine dodijeljena Nobelova nagrada upravo istraživačima koji su najviše doprinijeli razvoju metode elektronske krio-mikroskopije i doprinijeli rješavanju struktura makromolekula.

Od početka 2013. godine 50.28% lanaca iz skupa N907 pohranjeno je u bazu PDB, a od početka 2014. godine njih 40.68%. Tako, slično kao i u skupu N1212, oko polovice lanaca nije moglo biti zastupljeno u skupovima drugih autora objavljenih u ranijim godinama. Kod skupa N907 nisu uključene strukture određivane elektronskom krio-mikroskopijom. Uočava se rast broja struktura određenih do 2014., i pad broja struktura riješenih metodama rentgenske strukturne analize i NMR nakon te godine. Također, vidi se kako je u svim godinama udio struktura određenih metodom NMR u otopini mali, i ukupno gledano iznosi 12.57%.



Slika 33. Broj lanaca po godinama pohrane u bazu PDB i po eksperimentalnim metodama kojima je određena struktura u skupu N907.

U tablicama 28 i 29 dane su osnovne osobine skupova N1212 i N907. Uočava se veliki raspon TM segmenata u lancima, tj. od lanaca s jednim TM segmentom do onih s 24 TM segmenta, pa je skup N1212 jedini skup koji ima lance s 24 TM segmenta. Radi se o naponskom natrijumovu-kanalu (2 lanca) čija je struktura određena 2016. godine elektronskom krio-mikroskopijom s rezolucijom 3.6 Å i 3.8Å.

Tablica 28. Osnovne osobine skupa N1212 iskazane po lancu i za cijeli skup.

skup N1212	AK	TM segment	AK u membrani	% AK u membrani
minimum po lancu	17	1	6	2.12
maksimum po lancu	5037	24	548	88.24
srednja vrijednost po lancu	328.16	5.21	111.68	40.88
ukupno – skup	397729	6318	135362	

AK – aminokiselina, TM – transmembranski

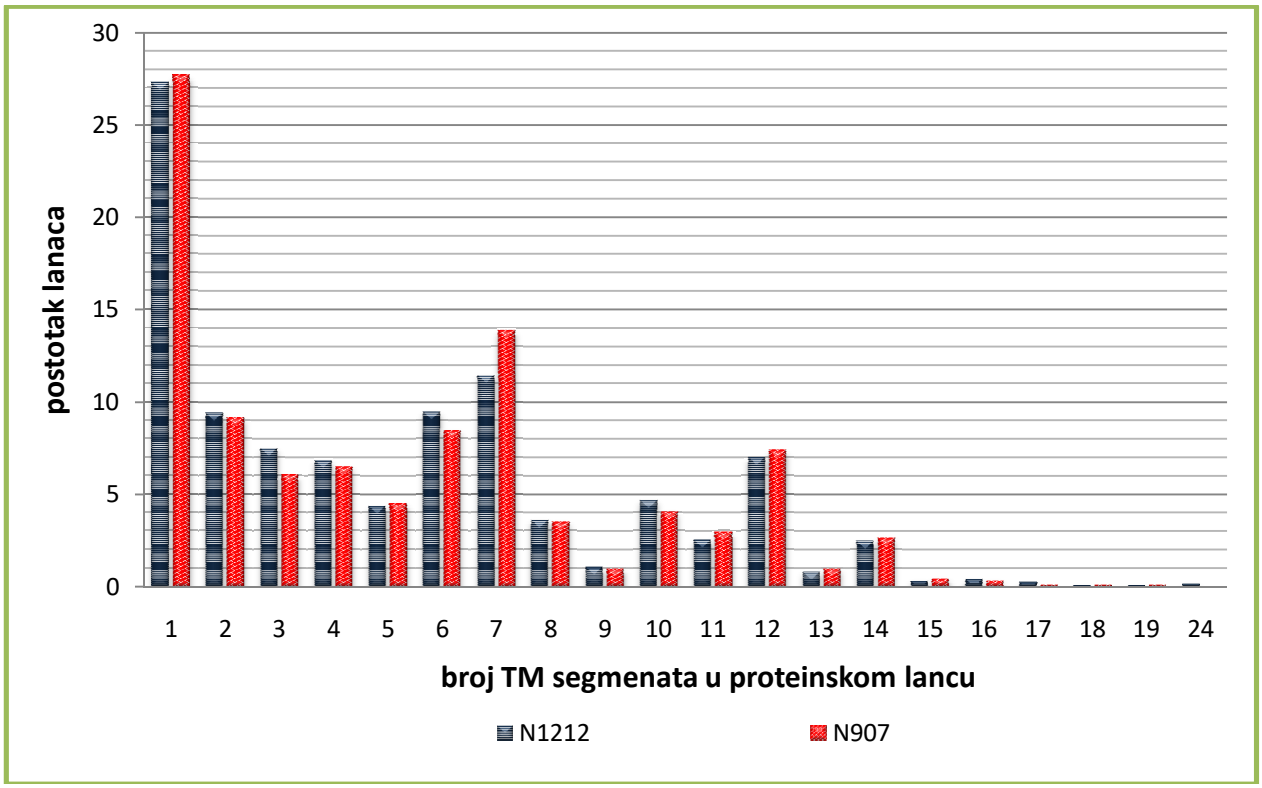
Tablica 29. Osnovne osobine skupa N907 iskazane po lancu i za cijeli skup.

skup N907	AK	TM segment	AK u membrani	% AK u membrani
minimum po lancu	22	1	6	2.12
maksimum po lancu	1321	19	443	84.00
srednja vrijednost po lancu	295.80	5.28	114.80	43.16
ukupno – skup	268294	4790	104127	

AK – aminokiselina, TM – transmembranski

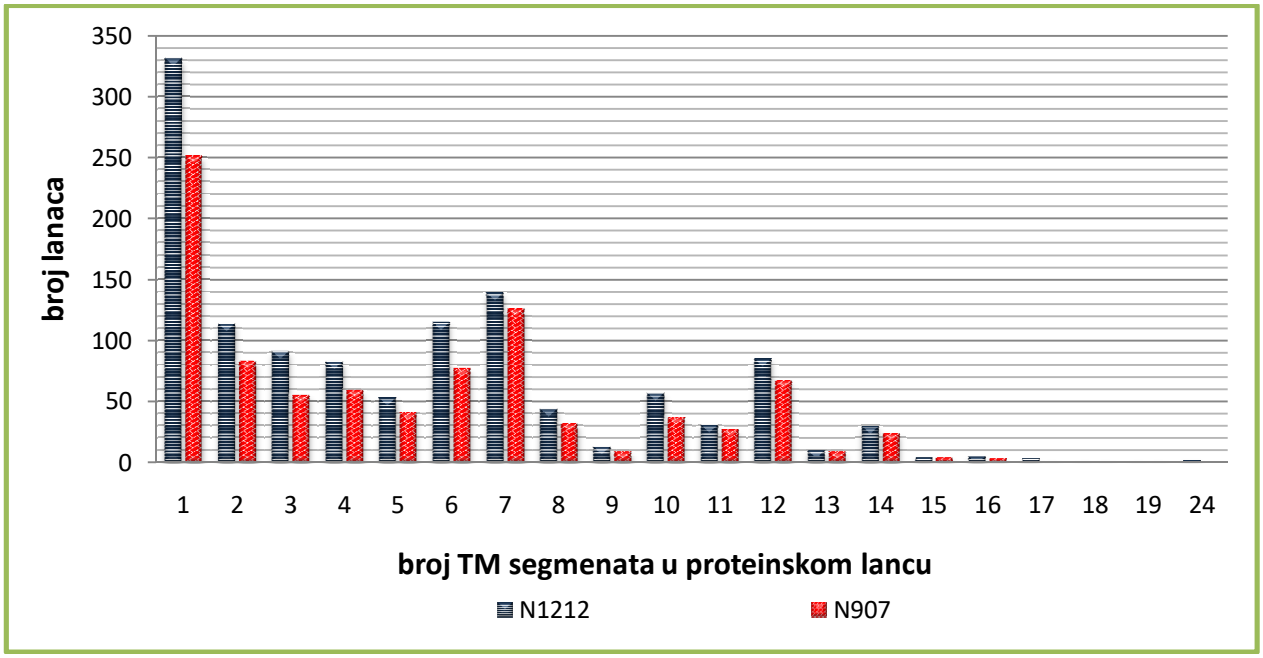
Nadalje, prosječni broj TM segmenata za skup N1212 i za skup N907 (kao i srednji postotak aminokiselina u membrani po proteinskom lancu) iznosi oko 5.25, i viši je nego u početnim skupovima drugih autora.

U raspodjeli postotka lanaca s određenim brojem TM segmenata za skupove N1212 i N907 uočava se ujednačenost među skupovima, kao i nešto veći postotak lanaca s 1, 2, 6, 7 i 12 TM segmenata u oba skupa (slika 34).



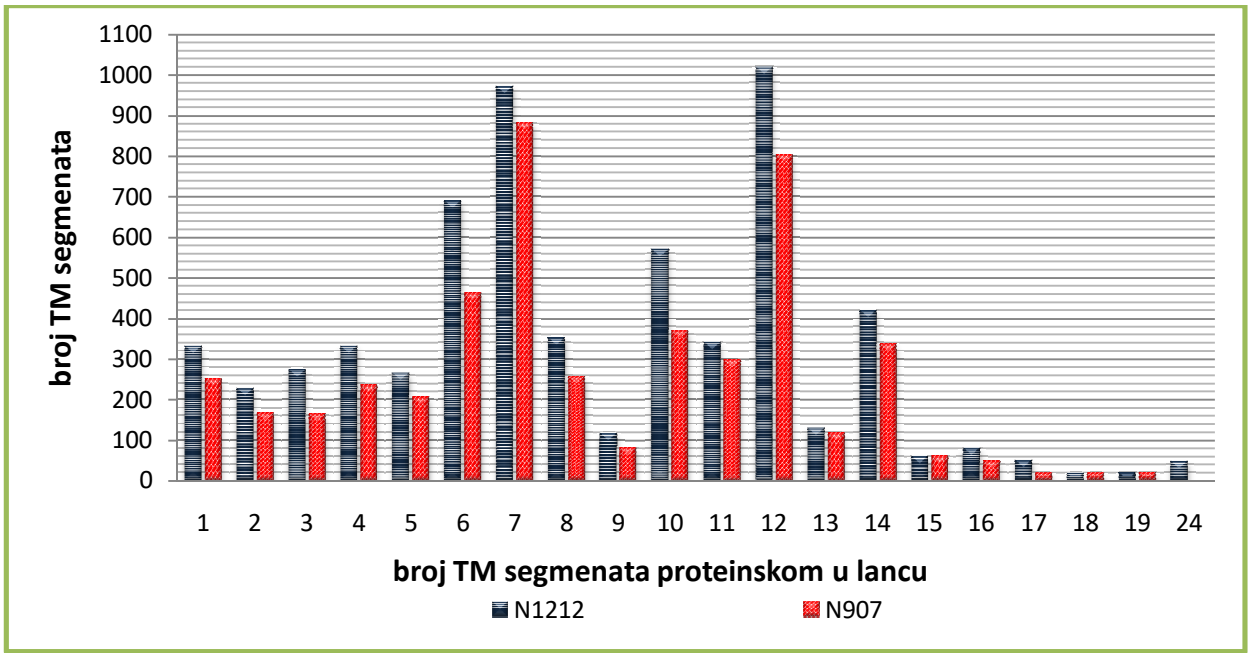
Slika 34. Postotak proteinskih lanaca u početnim skupovima N1212 i N907 po podskupovima lanaca istog broja TM segmenata.

Primjećuje se da je postotak lanaca sa 7 TM segmenata viši u skupu N907 nego u N1212, dok je za lance s 3 i 6 TM segmenata u istom skupu taj postotak niži za skup N907. Vidi se (na slici 35), da je najveći doprinos broju lanaca u skupu zapravo od broja lanaca s jednim TM segementom, a on je skoro dva puta veći od doprinosa podskupine lanaca sa 7 TM segmenata koja sljedi po doprinosu iza prve podskupine.



Slika 35. Broj proteinskih lanaca u početnim skupovima N1212 i N907 po podskupovima lanaca istog broja TM segmenata.

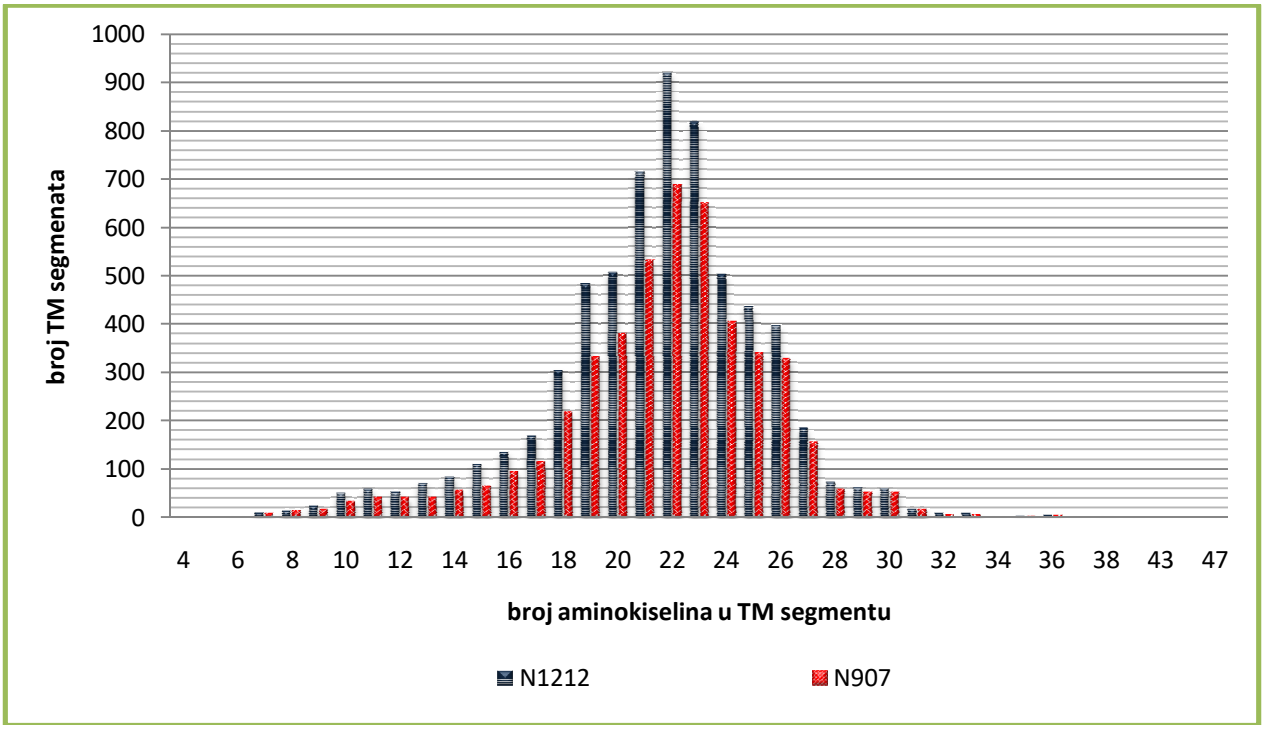
No, slike 34 i 35 ne daju pravu informaciju o važnosti tih podskupina lanaca u skupu u smislu broja TM segmenata koji donosi u ukupni skup svaka od podskupina, a ta je informacija zbirno dana na slici 36. Za razliku od slike 34 koja daje postotni udio lanaca u skupu po broju TM segmenata u lancu, slika 36 daje potpuno novi pogled na značajnost pojedinih skupina lanaca. Vidi se kako su skupine sa 7 i 12 TM segmenata u lancu najznačajnije u doprinosu ukupnom broja TM segmenata u skupovima N1212 i N907, a time i ukupnoj složenosti skupova.



Slika 36. Broj TM segmenata u skupovima N1212 i N907 po podskupovima lanaca istog broja TM segmenata.

Potom po doprinosu ukupnom broju TM segmenata slijede skupine sa 6, 10 i 14 TM segmenata. Zanimljivo je uočiti da je skupina s 12 TM segmenata u postotnom udjelu po broju lanaca peta po redu (slike 34 i 35), ali zbog velikog broja TM segmenata postaje najznačajnija podskupina lanaca u skupu N1212 (slika 36) i druga po značajnosti u skupu N907 po doprinosu ukupnom broju TM segmenata. Dodatno, vidi se u početnom skupu N1212 da je najzastupljenija skupina po broju lanaca (ona s jednim TM segmentom) po doprinosu ukupnom broju TM segmenata u izabranom skupu tek je na 9. mjestu. Slična situacija je i za skup N907, u kojem je podskupina lanaca s jednim TM segmentom tek na 8. mjestu po doprinosu ukupnom broju TM segmenata u izabranom skupu.

Osim analize značajnosti pojedine podskupine lanaca, važna je i analiza broja segmenata u ovisnosti o broju aminokiselina u segmentu (slika 37).



Slika 37. Raspodjela broja TM segmenata po duljinama TM segementa u skupovima N1212 i N907.

Vidi se da su najzastupljeniji segmenti duljina 21-23 aminokiseline, i samo te tri skupine segmenata imaju 2453 TM segmenata ili 38.83% ukupnog broja TM segmenata u skupu N1212. Iste skupine segmenata u skupu N907 imaju 1870 TM segmenata ili 39.04% ukupnog broja TM segmenata skupa. U oba skupa (N1212 i N907), skupine segmenata duljina 18-26 po svojoj brojnosti prelaze 80% ukupnog broja segmenata. U ukupnom broju segmenata u skupu N1212 preko 90% su segmenti koji imaju duljinu između 15 i 27 aminokiselina (koji su zastupljeni značajnije, tj. preko 100 puta). U intervalu između 10 i 30 aminokiselina taj je postotak 98.24%. Kraći segmenti obično su tipa zavoja (engl. *loop*) koji su u bazi OPM upisani kao TM segmenti. Ako se usporede raspodjele broja TM segmenata po duljinama za skupove N1212 i N907 vidi se pomak udesno raspodjele u skupu N907 (odnosno ka duljim segmentima).

3.3.4. Rezultati dobiveni primjenom algoritama 1, 2, 3 na skupove N1212 i N907

Algoritmi 1, 2 i 3 razvijeni u disertaciji i optimirani na drugim skupovima iz literature primijenjeni su u izboru najnovijih reprezentativnih skupova integralnih membranskih proteina alfa vrste polazeći od početnih skupova N1212 i N907. Izbor će se provesti uz prag identičnosti od 20% i prag sličnosti od 30%.

Analizom skupa na razini 20% identičnosti mogu se dodatno usporediti rezultati sa skupom "Grupe München", a rezultati izbora skupova uz prag sličnosti 30% usporedit će se s rezultatima "Grupe Sydney". Zbirna analiza broja lanaca u reprezentativnim skupovima dobivenim na razinama identičnosti 20% i sličnosti 30% uz primjenu razvijenih algoritama 1, 2 i 3 na početne skupove N1212 i njegov podskup N907, dana je u tablici 30.

Tablica 30. Vrijednosti parametara kvalitete reprezentativnih skupova dobivenih analizom skupova N907 i N1212 Algoritmima 1, 2 i 3 uz prag identičnosti 20% i sličnosti 30%.[a]

	lanaca	AK_{uk}	TM_{uk}	$Q_{2,rnd,sr}$	Bin_{uk}	$S_{0,uk}$	TM_{sr}	$S_{0,sr}$	$Q_{2,rnd}$
N1212	1212	397729	6318	0.587	213792	32175	5.21	26.55	0.55
N907	907	268294	4790	0.575	157681	23917	5.28	26.37	0.53
N907 20% ident.									
Algoritam 1	281	87240	1520	0.589	49212	7704	5.409	27.416	0.59
Algoritam 2	**281**	87696	**1528**	0.589	49490	**7744**	5.438	27.559	0.59
Algoritam 3	280	86932	1514	0.588	49087	7657	5.407	27.346	0.59
N1212 20% ident.									
Algoritam 1	**347**	121258	1836	0.605	60241	9448	5.291	27.228	0.61
Algoritam 2	339	119744	1827	0.606	59630	9426	5.389	27.805	0.61
Algoritam 3	345	121383	**1846**	0.605	60436	**9499**	5.351	27.533	0.60
N907 30% sličnosti									
Algoritam 1	**192**	60741	917	0.613	32428	4745	4.776	24.714	0.61
Algoritam 2	186	59657	920	0.61	31975	4751	4.946	25.543	0.61
Algoritam 3	184	60355	**942**	0.607	32779	**4879**	5.120	26.516	0.61
N1212 30% sličnosti									
Algoritam 1	234	85721	1110	0.629	39592	5883	4.744	25.141	0.63
Algoritam 2	227	83253	1116	0.619	39014	5873	4.916	25.872	0.62
Algoritam 3	**234**	86805	**1156**	0.624	40596	**6113**	4.940	26.124	0.62

[a] oznake kratica dane su ispod tablice 19

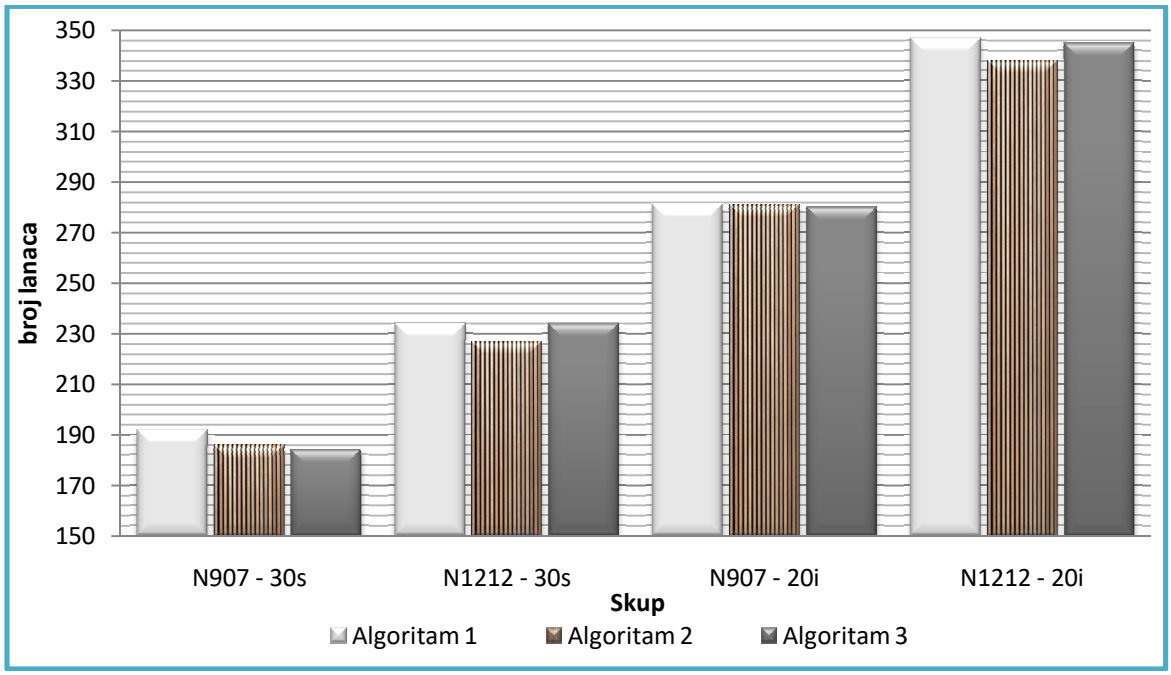
Vidi se da je razlika u broju lanaca dobivenih ovim algoritmima u rasponu od 5%, i da je najveći broj lanaca izdvojen Algoritmom 1. Najveći broj TM segmenata u reprezentativnom skupu izabire Algoritam 3, i rezultati su u rasponu od oko 4%, izuzev u slučaju primjene algoritama na razini 20% identičnosti na skupu N907 gdje je najviše segmenata u skupu dao Algoritam 2. Ovdje je vrijedno napomenuti da je vrijeme izvođenja Algoritma 3 u odnosu na Algoritme 1 i 2 znatno kraće.

Vrijednosti ukupnog entropijskog koeficijenta S_0 ($S_{0,uk}$) najveće su za skupove koji imaju najveći broj TM segmenata, unatoč tomu što ukupni broju lanaca nije nužno najveći. Tako, analizom početnog skupa N1212 Algoritmom 3 uz prag sličnosti 30% dobiven je najmanji broj lanaca (njih 184), dok je ukupna složenost skupa, a time i težina predviđanja strukture najveća za ovaj skup ($S_{0,uk}$ = 4879). Dakle, iako je broj lanaca manji za 4.35% u izabranom reprezentativnom skupu, ukupni entropijski koeficijent veći je za 2.82%, i (u grubo) toliko bi izabrani skup lanaca trebao biti teži za predviđanje. Ovaj rezultat zapravo upućuje na to kako Algoritam 3 u odnosu na druge metode pokazuje bolja svojstva u primijeni na većim skupovima. Dodatno, osjetno je bolji rezultat dobiven polazeći od skupa N1212 nego od skupa N907.

Međutim, Algoritam 3 osjetljiv je u radu na poredak lanca u početnom skupu, što se može smatrati i kao određena slabost. Dodatno, Algoritam 3 osjetljiv je s obzirom na broj međusobno sličnih lanaca u početnom skupu. To se, za definirani prag sličnosti (identičnosti), ogleda u

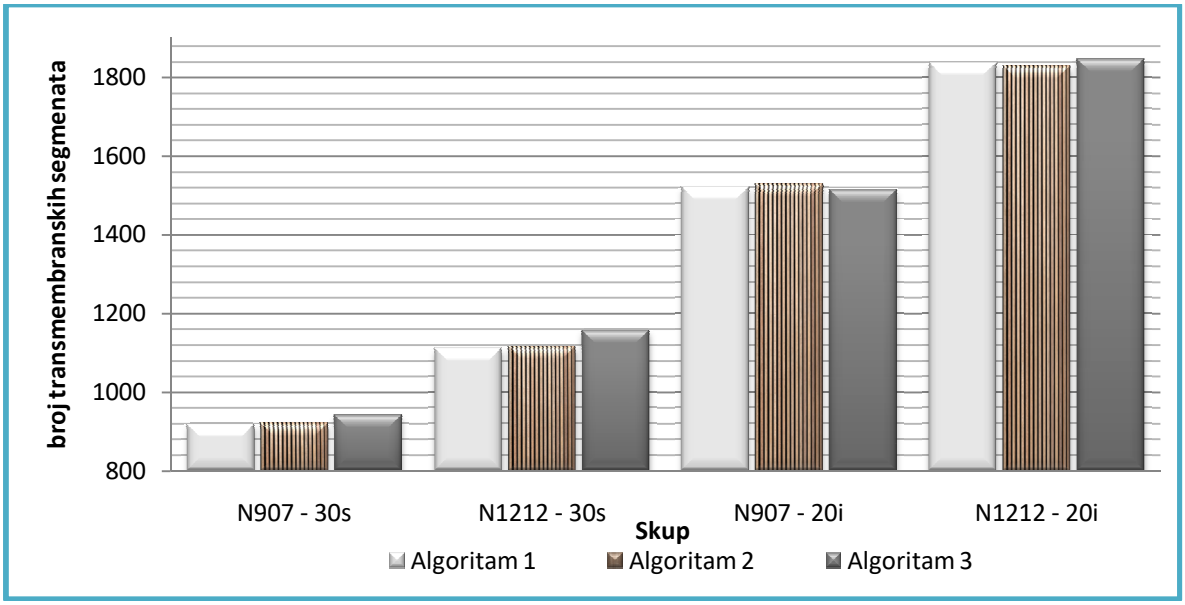
ulaznoj matrici sličnosti/identičnosti u većem broju vrijednosti koje nisu jednake nuli. Ta ograničenja Algoritma 3 zapravo su poticaj za njegovo unapređenje u narednom periodu.

Na slici 38 grafički je prikazan broj lanaca u reprezentativnim skupovima izabranim Algoritmima 1, 2 i 3 polazeći od skupova N1212 i N907, dok je na slici 39 prikazan broj TM segmenata u reprezentativnim skupovima.



Slika 38. Broj lanaca u reprezentativnim skupovima izabranim Algoritmima 1, 2 i 3 polazeći od skupova N1212 i N907.

Na slikama 38 i 39 „N907 – 30s“ znači da je početni skup N907, a reprezentativni skup izabran je uz prag sličnosti od 30% (označen kao '30s'). Analogno je s drugim oznakama za skup N1212 i npr. za prag identičnosti 20% (označen kao '20i').

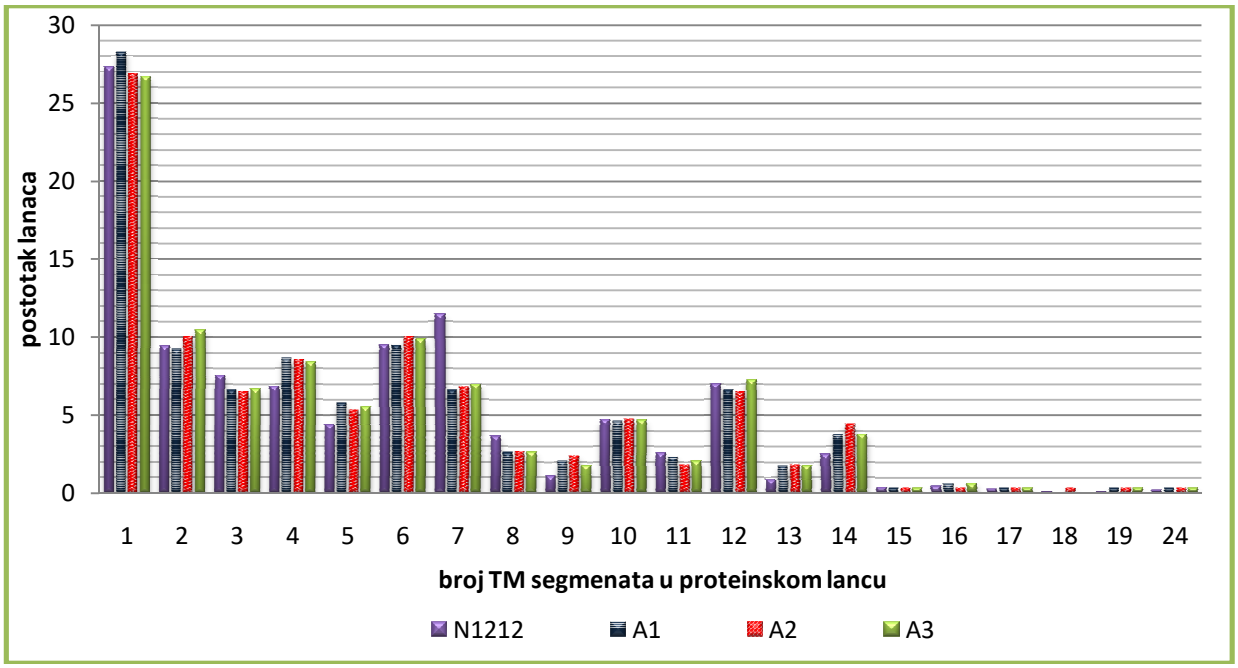


Slika 39. Broj transmembranskih segmenata u reprezentativnim skupovima izabranim Algoritmima 1, 2 i 3 polazeći od skupova N1212 i N907.

Na slici 38 pri izboru reprezentativnih skupova iz N1212 i N907 najbolje rezultate prema broju lanaca u izabranom skupu daje Algoritam 1. Međutim, po ukupnom broju TM segmenata u izabranom skupu nešto bolje rezultate dao je Algoritam 3 (slika 39).

3.3.4.1. Rezultati dobiveni na skupu N1212 uz prag identičnosti 20%

Raspodjela postotnog udjela lanaca u početnom skupu N1212 i u izabranim reprezentativnim skupovima dobivenim primjenom razvijenih algoritama na razini identičnosti 20% pokazuju značajnu razliku u postotku lanaca sa 7 TM segmenata (slika 40).

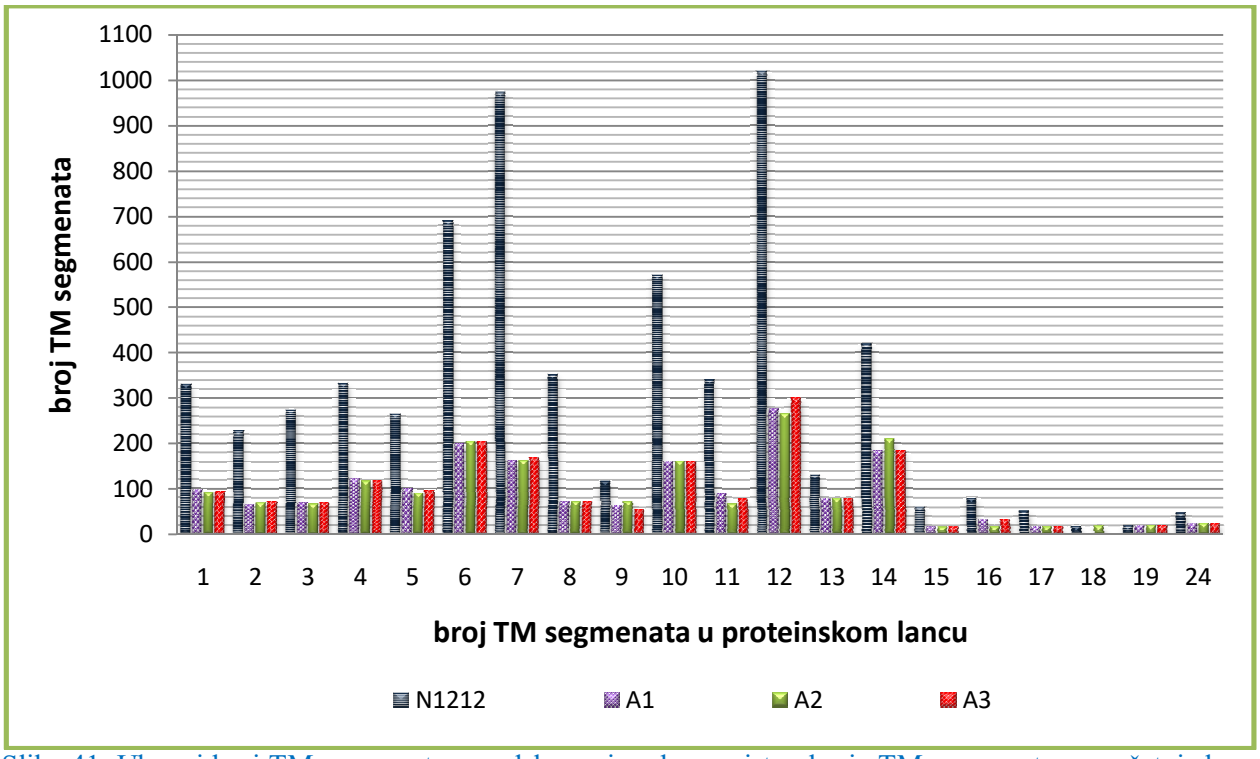


Slika 40. Raspodjela broja lanaca (u postotcima) po podskupovima lanaca istog broja TM segmenata u reprezentativnim skupovima izabranim Algoritama 1 (A1), 2 (A2) i 3 (A3) iz skupa N1212 (uz prag identičnosti 20%).

Smanjenje udjela takvih lanaca u reprezentativnom skupu izabranom Algoritmom 3 iznosi 4.51%, dok za skup izabran Algoritmom 2 to smanjenje iznosi 4.84%. To ukazuje na visoku međusobnu identičnost među lancima početnog podskupa lanaca koji imaju 7 TM segmenata iz skupa N1212. Stoga, u tom se podskupu dogodi značajnije smanjenje broja lanaca nakon primjene algoritama za redukciju zalihosti među proteinima. Kako se dogodio i postotni prirast u skupini lanaca s 14 TM segmenata, trebalo je dodatno ispitati kakva je sprega ovih podskupova lanca (sa 7 i 14 TM segmenata). To je provedeno programom EMBOSS [49], računajući njihove međusobne identičnosti/sličnosti. Međusobna identičnost svih lanaca sa 7 i 14 TM segmenata u prosjeku iznosi 10%, a ako se uzmu u obzir samo lanci koji pokazuju previsoku identičnost (iznad 20%), srednja vrijednost međusobne identičnosti iznosi samo 20.23%. Ovo nas vodi na zaključak kako lanci od 14 TM segmenata nisu značajno utjecali na postotno smanjenje lanaca sa 7 TM segmenata, nego njihova međusobna sličnost unutar podskupa. S druge pak strane, vidi se određeni blagi porast postotka lanaca s 4, 5, 9, 13 i 14 TM segmenata.

Na slici 41 dan je prikaz ukupnog broja TM segmenata razdvojen po podskupinama lanaca istog broja TM segmenata. Na primjer, za podskupinu lanaca s jednim TM segmentom zbrojeni su svi TM segmenti u svim lancima te podskupine, zatim je učinjeno isto za podskupinu lanaca s dva TM segmenta, itd. Vidi se da je doprinos podskupine lanaca s 12 TM segmenata u ukupnom broju TM segmenata u skupu ostao i dalje najveći (kao i u početnom skupu N1212). Međutim, doprinos skupine lanaca sa 7 TM segmenata (slika 41) pao je s drugog mjesta u skupu

N1212 (prema ukupnom broju TM segmenata – slika 41) na 4. mjesto u reprezentativnim skupovima i to u izborima svih algoritama. Potom, na drugom mjestu slijedi doprinos podskupine lanaca sa 6 TM segmenata u ukupnom broju TM segmenata u reprezentativnim skupovima izabranim sa sva tri algoritma.

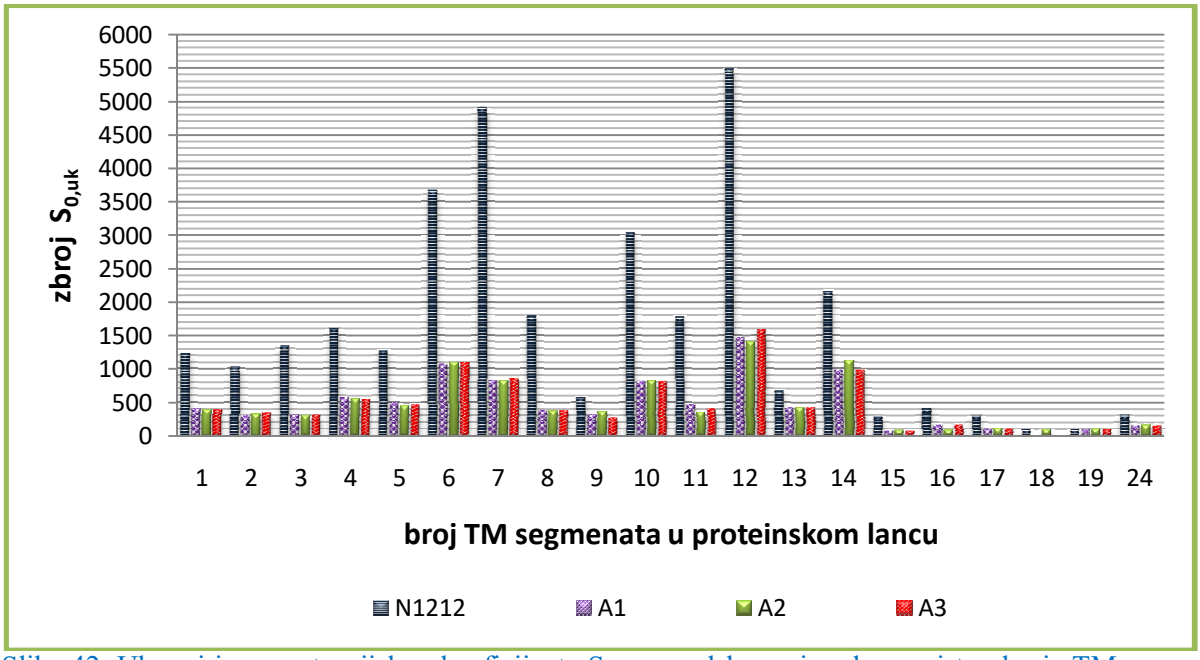


Slika 41. Ukupni broj TM segmenata u podskupovima lanaca istog broja TM segmenata za početni skup N1212 i skupove izabrane Algoritmima 1, 2 i 3 (za prag identičnosti 20%).

Značajnija promjena u doprinosu ukupnom broju TM segmenata u izabranom skupu uočava se za podskupinu s 14 TM segmenata koji je s 5. mjesta u početnom skupu N1212 došao na drugo/treće mjesto po doprinosu (iako je u postotnom udjelu prema broju lanaca slabo zastupljena). Pad doprinosa podskupine sa 7 TM segmenata ukupnom broju TM segmenata u reprezentativnim skupovima izabranim Algoritmima 1, 2 i 3 na slici 41 prati pad postotnog udjela broja lanaca u početnom skupu N1212 i u izabranim reprezentativnim skupovima (slika 40).

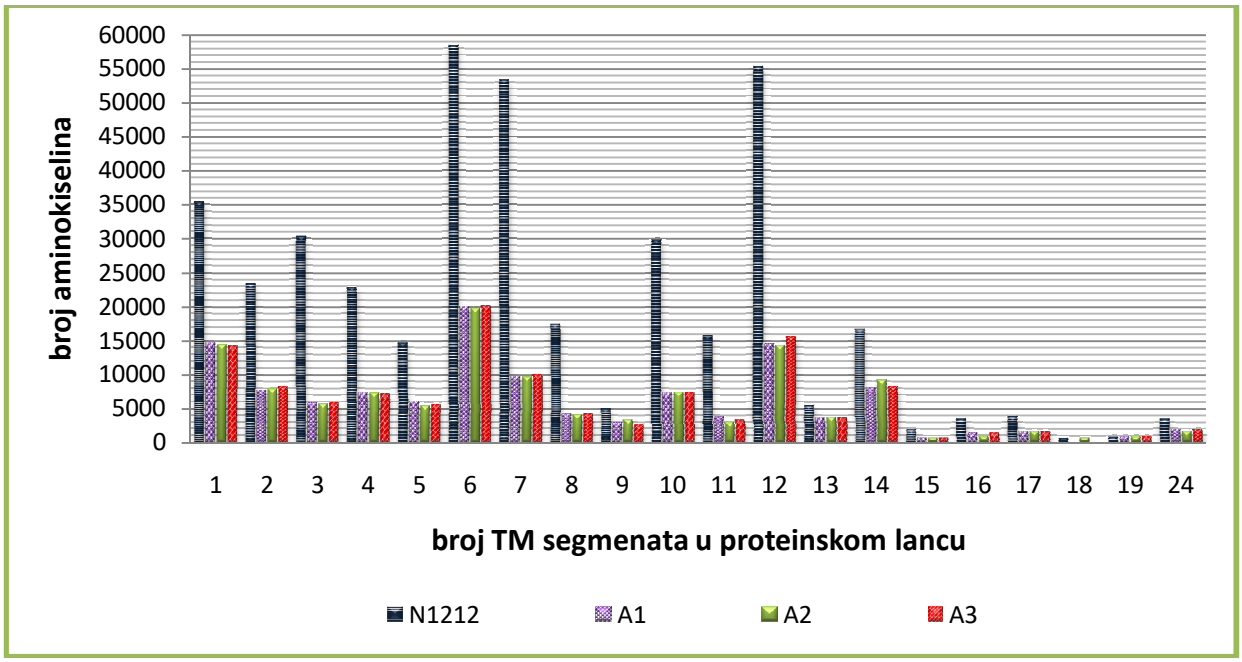
Izbori reprezentativnih skupova svakim od tri algoritma ne pokazuju (međusobno gledano) značajne razlike u doprinosu ukupnom broju TM segmenata cijelog skupa. Mali izuzetak je kod skupine s 12 TM segmenata kod koje odskače doprinos broju TM segmenata kod skupa izabranog Algoritmom 3, i kod skupine s 14 TM segmenata kod koje odskače izbor Algoritmom 2.

Usporedba ovisnosti (raspodjele) (*a*) zbroja TM segmenata (slika 41) i (*b*) zbroja doprinosa složenosti $S_0 = \ln(W_0)$ (slika 42) po lancima istog broj TM segmenata pokazuje značajno slaganje, s korelacijskim koeficijentima 0.9964 (skup N1212), te 0.9963, 0.997 i 0.997. za reprezentativne skupove izabrane (redom) Algoritmima 1, 2 i 3. Ovo značajno slaganje potvrđuje opravdanost uporabe broja TM segmenata kao jednostavnog i korisnog kriterija pri izboru lanaca algoritmima razvijenim u disertaciji.



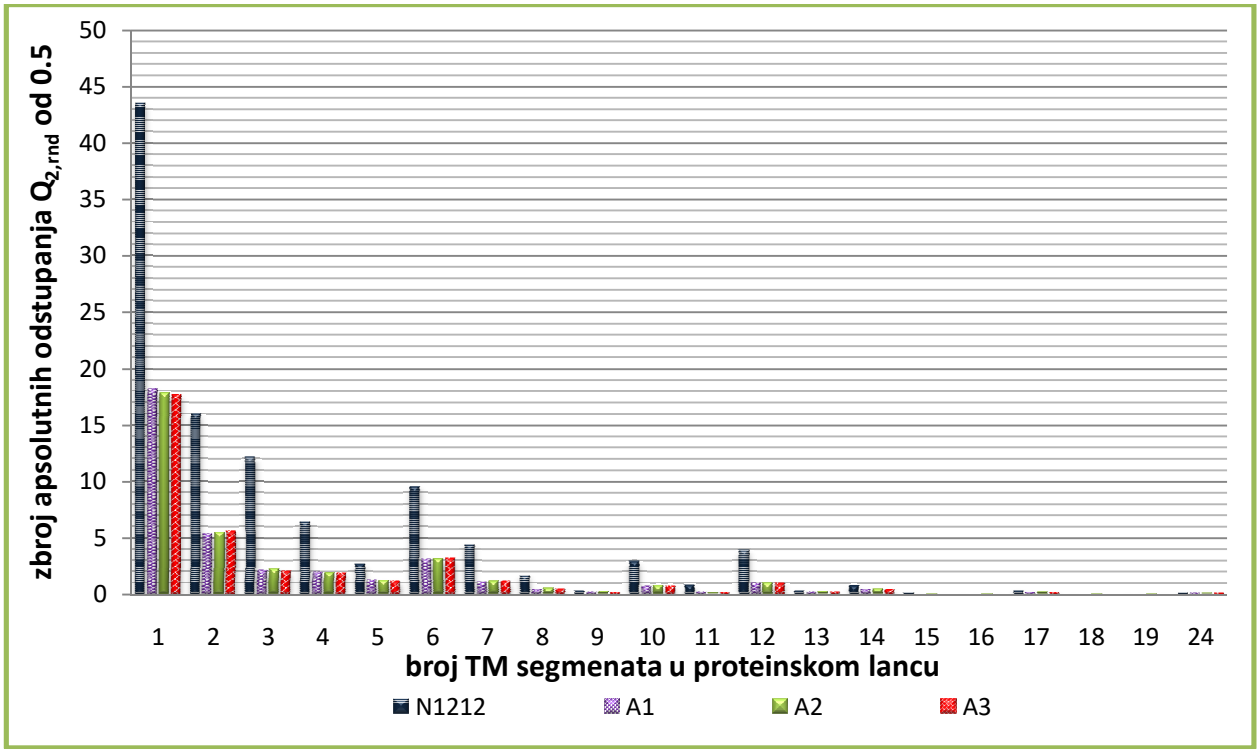
Slika 42. Ukupni iznos entropijskog koeficijenta $S_{0,uk}$ u podskupovima lanaca istog broja TM segmenata za početni skup N1212 i skupove izabrane Algoritmima 1, 2 i 3 (za prag identičnosti 20%).

Ujedno, to pokazuje da je duljina slijedova manje značajan kriterij kvalitete i pokazatelj složenosti strukture lanaca u odnosu na broj TM segmenata, što potvrđuje našu početnu istraživačku pretpostavku. Ovo potkrepljuje i usporedba raspodjele ukupne složenosti $S_0 = \ln(W_0)$ (slika 43) i raspodjele ukupnih brojeva aminokiselina (slika 44) po skupinama lanaca s jednakim brojem TM segmenata. Naime, vidi se da podskupine lanaca s 1, 2, 3 i osobito sa 6 TM segmenata doprinose bitno više ukupnim brojevima aminokiselina u reprezentativnim skupovima nego ukupnim složenostima struktura. To vrijedi podjednako za skupove izabrane sa sva tri algoritma.



Slika 43. Ukupni broj aminokiselina u podskupovima lanaca istog broja TM segmenata za početni skup N1212 i skupove izabrane Algoritmima 1, 2 i 3 (za prag identičnosti 20%).

Dodatno se uočava da je doprinos podskupina lanaca s 1 i 12 TM segmenata ukupnim brojevima aminokiselina u izabranim skupovima podjednak, dok prva podskupina lanaca doprinosi tri puta manje ukupnoj složenosti reprezentativnih skupova nego druga. Prosječna duljna lanaca sa 6 TM segmenata znatno je veća u reprezentativnom skupu (590 aminokiselina) nego u početnom skupu N1212 (508 aminokiselina). To znači da algoritmi uz povećanje broja TM segmenata u reprezentativnim skupovima, između dva lanca s istim brojem TM segmenata biraju lance veće duljine. Na taj se način povećava složenost skupa, tj. povećava se $S_0 = \ln(W_0)$.



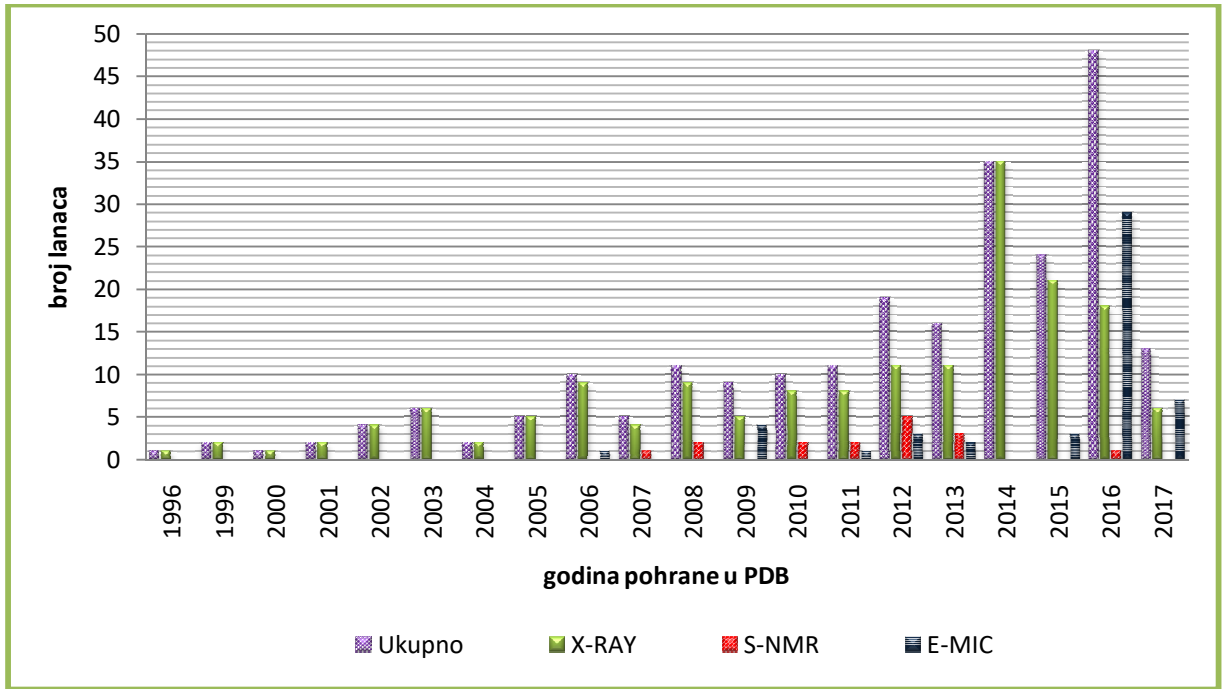
Slika 44. Ukupni iznos koeficijenta ($Q_{2,rnd} - 0.5$) u podskupovima lanaca istog broja TM segmenata za početni skup N1212 i skupove izabrane Algoritmima 1, 2 i 3 (za prag identičnosti 20%).

Rezultati analize na parametru $Q_{2,rnd}$, odnosno njegovim srednjim apsolutnim odstupanjima od 0.5 zbrojeno po lancima istog broja TM segmenata prikazani su na slici 44. Vrijednost $Q_{2,rnd} = 0.5$ najmanja je moguća točnost uravnoteženog nasumičnog modela u predviđanju dva stanja sekundarne strukture. Promatrano po ovom parametru, teži za predviđanje strukture oni su lanci koji imaju $Q_{2,rnd} - 0.5$ bliži vrijednosti 0. Na toj slici vidi se da je daleko najveći doprinos vrijednosti $(Q_{2,rnd} - 0.5)$ cijelom skupu dolazi od lanaca s 1 TM segmentom, a u doprinosu slijede podskupine lanaca s 2, 6, 3, 4, 5 i 7 TM segmenata. Takvo se ponašanje pokazuje u sva tri izabrana reprezentativna skupa (sa svakim od tri algoritma). Prosječni broj aminokiselina po lancu za podskupinu lanaca s 1 TM segmentom u početnom skupu N1212 iznosi 106, a u reprezentativnim skupovima, preko 150. Kako su duljine TM segmenta (u toj podskupini svi lanci imaju 1 TM segment) u prosjeku 21 aminokiseline, to znači da će parametru $Q_{2,rnd}$ sve više odstupati od vrijednosti 0.5, koja odgovara idealnom (najtežem) nasumičnom modelu. Slično razmatranje vrijedi i za podskupine proteina s više TM segmenata, samo, kod tih lanaca, povećanje prosječne duljine slijeda prati i povećanje broja TM segmenata. To u konačnici povećava udio aminokiselina u membrani (u sekundarnoj strukturi α), pa je $Q_{2,rnd}$ bliži vrijednosti 0.5.

3.3.4.2. Rezultati dobiveni na skupu N1212 za prag sličnosti 30%

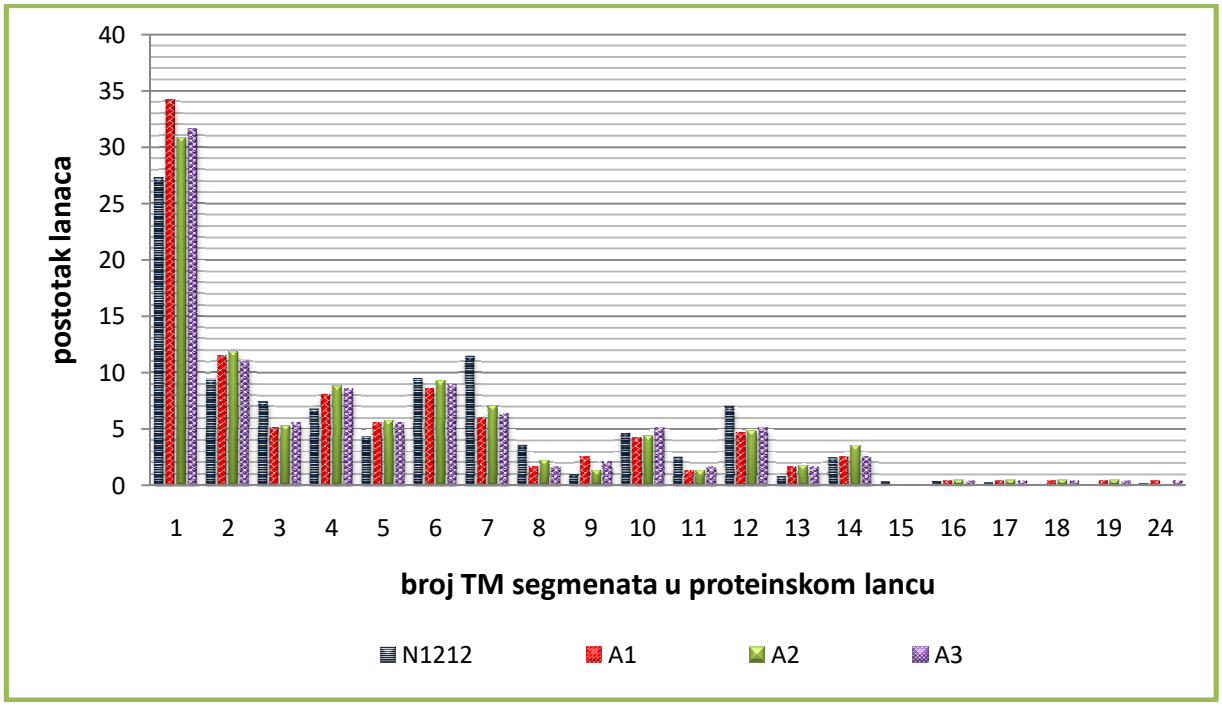
Reprezentativni skup za prag sličnosti 30% dobiven Algoritmom 3 iz početnog skupa N1212 sadrži proteinske lance čija je struktura određena s tri eksperimentalne metode i to: rendgenskom difrakcijom (168 lanaca), nuklearnom magnetskom rezonancijom (u otopini) (16 lanaca) i elektronskom mikroskopijom (50 lanaca). Ovaj reprezentativni skup (N234) ukupno sadrži 234 proteinska lanca od kojih je njih 58.12% pohranjeno u bazu PDB od početka 2013. godine, odnosno njih 51.28% od početka 2014. godine. Može se vidjeti na slici 45 da je skoro dvije trećine lanaca izabranih u skup N234 pohranjeno nakon izbora reprezentativnih skupova drugih autora iz disertacije [41,42,43]. Uočava se u ovom skupu da je u 2016. godini prvi put najveći broj struktura membranskih proteina eksperimentalno određen metodom elektronske mikroskopije (98 lanaca) a manje je lanaca određeno metoda rendgenske difrakcije (89 lanaca). K tome, od 98 lanaca čije su strukture određene elektronskom mikroskopijom u 2016. (i 16 lanaca u 2017.) godini (slika 32), u reprezentativni skup N234 izabrano je (redno) 29 odnosno 7 lanaca.

Taj je skup dan u Prilogu B (primarne strukture, oznake položaja TM segmenata, naziv, identifikacijski kod proteina u bazi PDB).



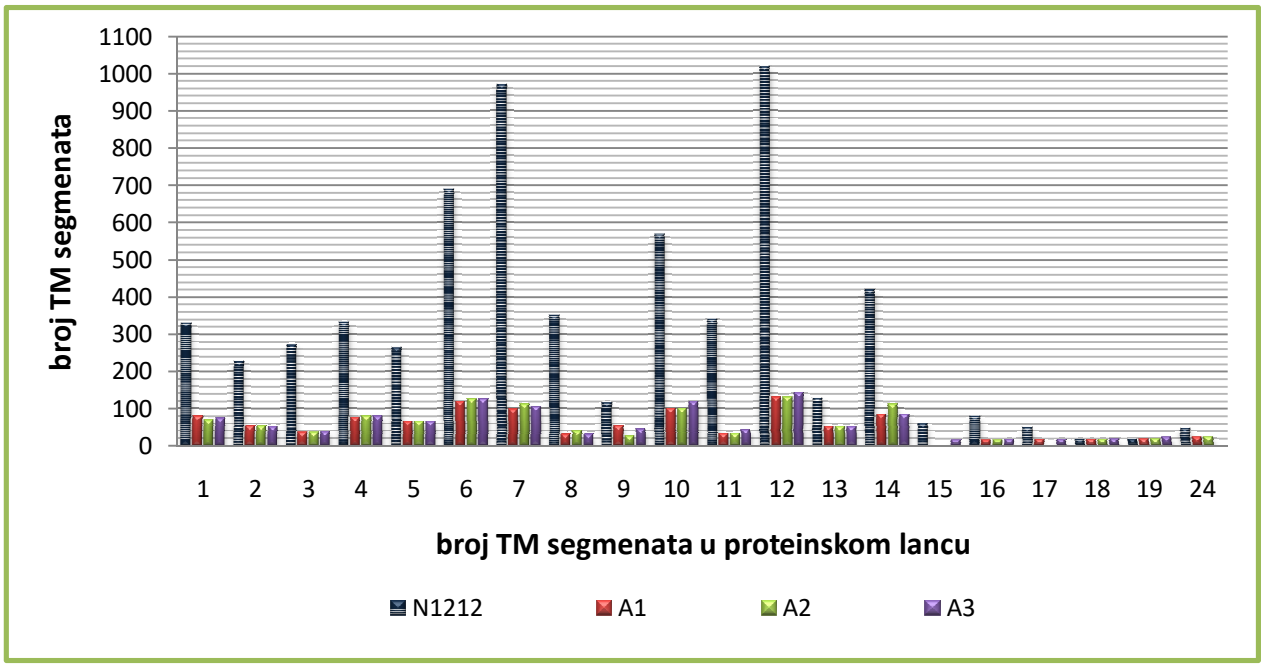
Slika 45. Broj lanaca po godinama pohrane u bazu PDB i po eksperimentalnim metodama kojima je određena struktura u reprezentativnom skupu N234 izabranom iz skupa N1212 (uz prag sličnosti 30%).

Analizirajući skup N234 vidi se značajniji porast postotnog udjela lanaca s jednim TM segmentom i to 6.88%, 3.53% i 4.31% u skupovima izabranim (redom) Algoritmima 1, 2 i 3 (slika 46). Nadalje, porastao je i postotni udio lanaca u podskupinama lanaca s 2, 4, 5, 9 i 13 TM segmenata. Također, značajniji pad u postotnoj zastupljenosti imaju podskupine lanaca sa 7 TM segmenata (slično kao i u slučaju izbora uz prag identičnosti 20%), te podskupine lanaca s 3, 8, 11 i 12 TM segmenata. Ovi podaci ukazuju na to da je među lancima podskupina s većim brojem TM segmenata u početnom skupu N1212 prisutna visoka međusobna sličnost. Stoga, provedba postupka izbora repezentativnih skupova izbacuje suvišne lance (one koji su previše slične drugim lancima) i tako smanjuje zalihost u tim podskupinama lanaca s više (3, 6, 7, 8, 11 i 12) TM segmenata.



Slika 46. Raspodjela broja lanaca (u postotcima) po podskupovima lanaca istog broja TM segmenata u reprezentativnim skupovima izabranim Algoritama 1 (A1), 2 (A2) i 3 (A3) iz skupa N1212 (uz prag sličnosti 30%).

Ukupan doprinos broju TM segmenata u skupu N1212 (i u reprezentativnim skupovima) pojedinih podskupina lanaca s istim brojem TM segmenata, prikazan je na slici 47. Najveći doprinos broju lanaca u izabranim skupovima dolazi (redom) od podskupina lanaca s 12, 6, 10, 7, 14, 4 TM segmenta. Najveće smanjenje doprinosa u odnosu na početni skup N1212 je kod podskupina s 12, 7, 6, 10 i 14 TM segmenata.



Slika 47. Ukupni broj TM segmenata u podskupovima lanaca istog broja TM segmenata za početni skup N1212 i skupove izabrane Algoritmima 1, 2 i 3 (za prag sličnosti 30%).

Kao i u slučaju prethodno analiziranoga izbora po identičnosti, raspodjelu ukupnog broja TM segmenata po podskupovima lanaca istog broja TM segmenata prati raspodjela po doprinosu složenosti strukture $S_0 = \ln(W_0)$ (entropijski doprinos). Ta je slika dana u Prilogu A. Dodatno, uočava se sličnost (a) raspodjela doprinosa ukupnom broju aminokiselina i (b) doprinosa ukupnom koeficijentu nasumične točnosti uravnoteženog nasumičnog modela za dvije sekundarne strukture $Q_{2,rnd}$, odnosno $(Q_{2,rnd} - 0.5)$. Ove slike (raspodjele) također su dane u Prilogu A.

3.3.4.3. Rezultati dobiveni na skupu N907 uz prag identičnosti 20% i prag sličnosti 30%

U tablici 30 dani su numerički parametri izabranih reprezentativnih skupova na razini identičnosti 20% polazeći od početnoga skupa membranskih proteina N907. Taj skup sadrži nešto točnije strukture membranskih proteina. Tako sva tri algoritma razvijena u disertaciji izabiru reprezentativne skupove sličnih osobina, a ukupno najbolji rezultat dobiven je Algoritmom 2. Taj reprezentativni skup sadrži 281 lanac, 1528 TM segmenata i uključuje proteinske lance membranskih proteina alfa vrste s najvišom ukupnom složenošću struktura (entropijski koeficijent) $S_{0,uk} = 7744$.

U slučaju izbora uz prag sličnosti 30%, kao ukupno najbolji reprezentativni skup može se smatrati skup sa 184 proteinska lanca dobiven Algoritmom 3. Taj skup ima 942 TM segmenta, i ukupnu složenost struktura (entropijski koeficijent) $S_{0,uk} = 4879$.

Kvalitativni izgled raspodjela osnovnih svojstava reprezentativnih skupova dobivenih Algoritmima 1, 2 i 3 polazeći od početnoga skupa N907 uz prag identičnosti 20% vrlo su slični onima dobivenim za početni skup N1212 (slike 39 do 43). To se odnosi na ovisnosti (raspodjele) po skupinama lanaca s istim brojem TM segmenata za sljedeća svojstva:

a) postotni udio broja proteinskih lanaca (slika 1.B),
b) doprinos ukupnom broju TM segmenata u skupu (slika 2.B),
c) doprinos ukupnom entropijskom koeficijentu u skupu $S_{0,uk}$ (slika 3.B),
d) doprinos ukupnom broju aminokiselina u skupu (slika 4.B),
e) doprinos ukupnom iznosu koeficijenta točnosti uravnoteženog nasumičnog modela s dva stanja sekundarne strukture $(Q_{2,rnd} - 0.5)$u skupu (slika 5.B).

Za slučaj izbora skupa uz prag sličnosti 30%, odgovarajuće slike koje odgovaraju onima pod a) do e), nalaze se u Prilogu B u slikama 6.B do slike 10.B

3.4. Zbirna sporedba rezultata na skupovima

Kako bi se uvidjele prednosti rada ovih algoritama poželjno je usporediti prosječnu vrijednost TM segmenata po lancu u reprezentativnim skupovima. Taj podatak može biti pokazatelj složenosti strukture proteina s obzirom da je uočena visoka korelacija između složenosti strukture i broja TM segmenata u proteinskim slijedovima na većim skupovima (slike 25 i 36). Razlike između prosječnih brojeva TM segmenata po proteinskom lancu mogu se analizirati na temelju podataka danih u tablici 31. Početni skupovi drugih autora M187, M1101, S481 i S392 čine prvu skupinu, i imaju prosječno 4.70 – 4.90 TM segmenata po lancu. Početni skupovi izdvojeni iz proteinskih baza u sklopu izrade disertacije čine drugu skupinu u tablici 31 i imaju oko 10% veći prosječni broj TM segmenata po lancu: N1212 (5.21) i N907 (5.28).

Reprezentativni skupovi izabrani iz početnih skupova drugih autora (M187, M1101, S481 i S392) imaju veći raspon prosječnog broja TM segmenata po lancu, i to od 2.25 (izbor algoritmom UniqueProt iz početnog skupa M1087) do 5.10 (izbor Algoritmom 2 iz početnog skupa S392). Prosječni broj TM segmenata po lancu za 10 reprezentativnih skupova (prvi deset

vrijednosti prosječnog broja TM segmenata po lancu u reprezentativnim skupovima iz tablice 31, 4.stupac) iznosi 4.45. U izboru reprezentativnih skupova Algoritmima 1, 2 i 3 iz početnih skupova N1212 i N907 izdvojenih u disertaciji (druga skupina u tablici 31), od 12 reprezentativnih skupova (prosječne vrijednosti 5.15 TM segmenata po lancu) čak njih 7 ima više od 5.1 TM segmenata po lancu. Taj broj (5.1) najveći je prosječni broj TM segmenata u reprezentativnim skupovima izabranim iz početnih skupova M187, M1101, S481 i S392 u prvom dijelu tablice.

Tablica 31. Usporedba početnih i reprezentativnih skupova s obzirom na prosječni broj TM segmenata u proteinskom lancu.

početni skup	prosječni broj TM segmenata po lancu	reprezentativni skup (Algoritam-prag identičnosti ili sličnosti)	prosječni broj TM segmenata po lancu
M1087	4.70	M166 (UniqueProt-20% identičnosti)	2.25
M1101	---	M190 (UniqueProt-20% identičnosti)	2.99
S481	4.86	S164 (Hobohm 2-20% identičnosti) N169 (A3-20% identičnosti)	5.00 5.03
S481	4.86	S121 (Hobohm 2-30% sličnosti) N127 (A3-30% sličnosti)	4.79 4.89
S392	4.90	S134 (Hobohm 2-20% identičnosti) N149 (A2-20% identičnosti)	4.76 5.10
S392	4.90	S101 (Hobohm 2-30% sličnosti) N115 (A2-30% sličnosti)	4.78 4.90
N1212	5.21	N347_A1 (20% identičnosti) N339_A2 (20% identičnosti) N345_A3 (20% identičnosti) N234_A1 (30% sličnosti) N227_A2 (30% sličnosti) N234_A3 (30% sličnosti)	5.29 5.39 5.35 4.74 4.92 4.94
N907	5.28	N281_A1 (20% identičnosti) N281_A2 (20% identičnosti) N280_A3 (20% identičnosti) N192_A1 (30% sličnosti) N186_A2 (30% sličnosti) N184_A3 (30% sličnosti)	5.41 5.44 5.41 4.78 4.98 5.12

Iz ukupne usporedbe za sve prethodno prikazane početne skupove i njima pripadajuće reprezentativne skupove, može se vidjeti da razvijeni Algoritmi 1, 2 i 3 izdvajaju u prosjeku više TM segmenata u reprezentativnim skupovima. Dodatno, oni izdvajaju i više lanaca, izuzev u samo jednom slučaju. Zbog korelacije između broja TM segmenata i ukupne konformacijske entropije modelne strukture proteina, proizlazi da Algoritmi 1, 2 i 3 razvijeni u disertaciji izdvajaju i složenije skupove. K tome, Algoritmi 1, 2 i 3 brže provode izbor reprezentativnih skupova (često i znatno brže). To im je dodatna prednost u odnosu na postojeće algoritme drugih autora koja će dolaziti više do izražaja kada se u njegovim primjenama bude polazilo od većih početnih skupova. Posebno je potrebno naglasiti osobine Algoritma 3 koji je pokazao bolje rezultate kada se u izboru polazilo od složenijih i većih početnih skupova. Nadalje, Algoritam 3 izbor reprezentativnog skupa lanaca provodi znatno brže, uz originalni način rada.

Usporedbom početnih skupova izdvojenih iz baza OPM[29] i PDB [20] u različitim vremenskim trenutcima (na različite datume), može se identificirati nove riješene strukture (koje

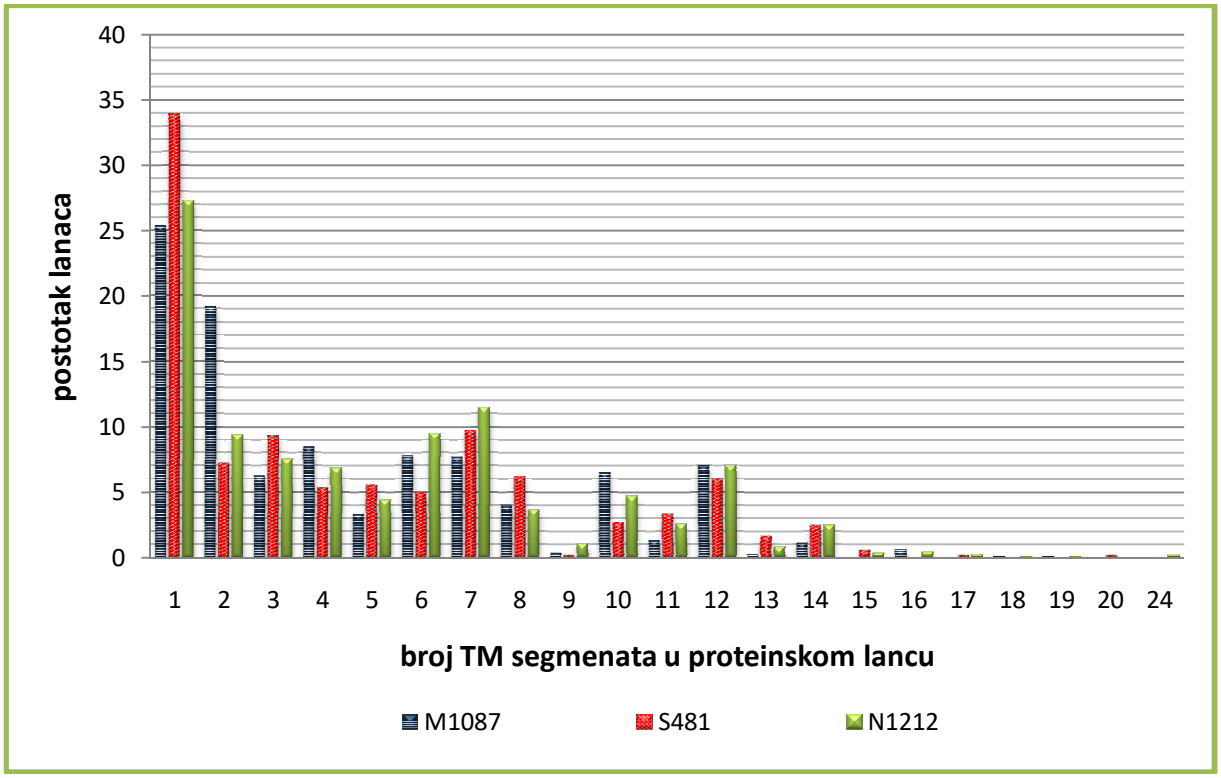
su se pojavile u novijim inačicama baza). Tako se mogu odvojiti razlike u izabranim skupovima nastale zbog povećanja broja lanaca s riješenim strukturama od razlika nastalih zbog različite kvalitete primijenjenih algoritama. Jedan dio razlike ima svoj uzrok u različitom postupku filtriranja proteinskih lanaca pri čemu se npr. najčešće izdvajaju (1) identični ili skoro identični lanci i izuzimaju iz početnog skupa ili (2) lanci čija je struktura određena s nižom točnošću (višom rezolucijom). Neki su autori proveli strožiji postupak filtriranja pri izboru početnih skupova, i zbog toga broj lanaca u njihovom početnom skupu bude manji. U nastavku analiziramo obilježja početnih skupova S481, M1087 i početnog skupa N1212 iz disertacije po podskupovima lanaca istog broja TM segmanata. Rezultati su dani u tablici 32, a skupovi su uspoređeni s obzirom na broj lanaca, ukupni broj aminokiselina i prosječne duljine lanaca. Zanimljivo je primijetiti izrazito veliku prosječnu duljinu lanaca s 1 TM segmentom u početnom skupu M1087 koja je (dodatno) veća od prosječne duljine lanaca s 2 TM segmenta u istom skupu. Također, vidi se velika prosječna duljina lanaca sa 3 i 6 TM segmenta u početnom skupu N1212. Nadalje, između podskupina lanaca s 9 TM i s 10 TM segmenata vidi se veliki skok u prosječnoj duljini lanaca, a zatim do podskupine s 13 TM segmenata dolazi do postupnog pada prosječne duljine.

Tablica 32. Usporedba glavnih obilježja proteinskih lanaca u skupovima S481, M1087 i N1212 po podskupovima lanaca istog broja TM segmenata.[a]

# TM segmenata u podskupu	S481			M1087			N1212		
TM	*lanaca*	AK_{uk}	AK_{sr}	*lanaca*	AK_{uk}	AK_{sr}	*lanaca*	AK_{uk}	AK_{sr}
1	163	20850	128	276	63546	**230**	331	35362	107
2	35	7122	203	**209**	40025	192	114	23343	205
3	45	8396	187	68	14248	210	**91**	30269	**333**
4	26	7452	287	92	29744	323	83	22846	275
5	27	8599	318	37	10291	278	**53**	14674	277
6	24	9085	379	85	28971	341	**115**	58469	**508**
7	47	15418	328	84	26981	321	**139**	53358	**384**
8	30	9615	321	44	15080	343	44	17376	395
9	1	514	**514**	4	1424	356	**13**	4936	380
10	13	9533	**733**	**71**	46022	648	57	29870	524
11	16	9191	574	15	8519	568	**31**	15723	507
12	29	16509	569	77	48202	626	**85**	55319	651
13	8	4272	534	3	2595	865	10	5405	541
14	12	6598	550	13	6443	496	**30**	16674	556
15	3	1518	506	0	0	0	4	1974	494
16	0	0	0	7	5042	720	5	3480	696
17	1	485	485	0	0	0	3	3778	1259
18	0	0	0	1	613	613	1	613	613
19	0	0	0	1	791	791	1	791	791
20	1	613	613	0	0	0	0	0	0
24	0	0	0	0	0	0	2	3469	1735
Ukupno	481	135770		1087	348537		1212	397729	

[a] # lan. – broj lanaca; AK_{uk} – ukupni broj aminokiselina; AK_{sr}- prosječna duljina lanca

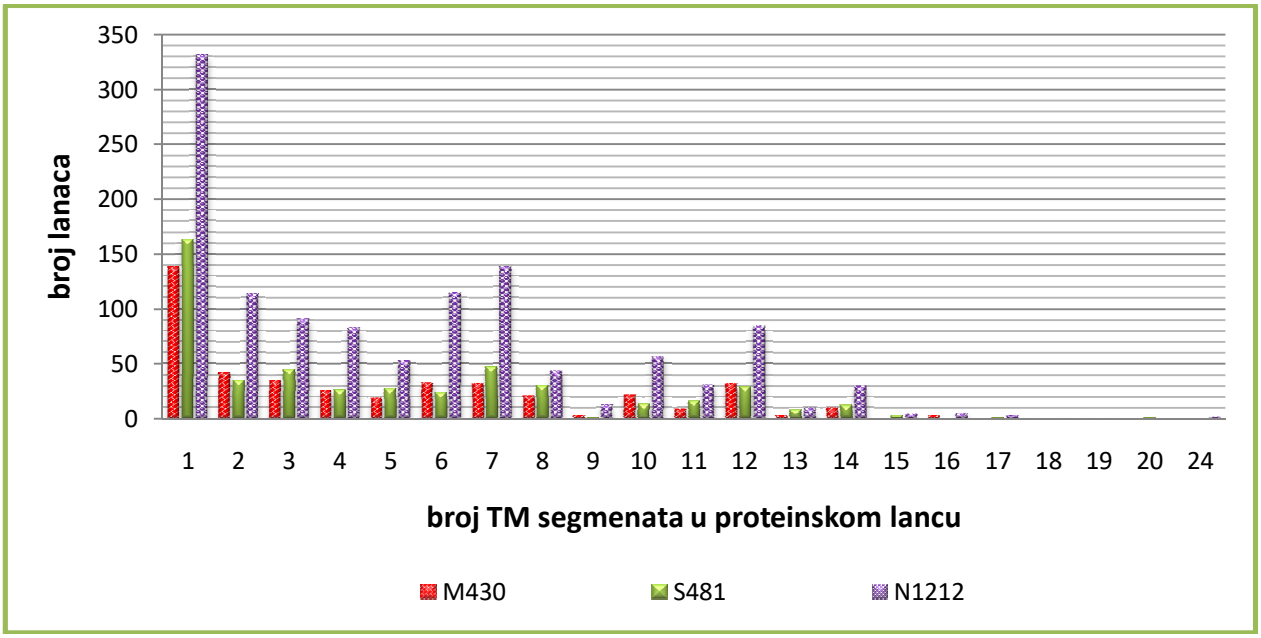
Na slici 48. vidi se kako su u odabranom skupu N1212 najzastupljeniji podskupovi lanaca (poredano od većih zastupljenosti) s 1, 7, 6, 2, 3, 12, 4 i 10 TM segmenata. Kod skupa M1087 taj poredak je: 1, 2, 4, 6, 7, 12 i 10, dok za skup S481: 1, 7, 3, 2, 8, 12, 5 i 4. Podskup lanaca s 1 TM segmentom najzastupljeniji je u svim skupovima, dok je u skupu M1087 na drugom mjestu podskup lanaca s 2 TM segmenta (u druga dva to je podskup sa 7 TM segmenata). Ovo pokazuje zašto u skupu M1087 (u prosjeku) ima malo TM segmenata po lancu, kako u početnom, tako i u reprezentativnim skupovima dobivenim primjenom algoritama na skup M1087 kao početni. Zanimljivo je primijetiti brojnost (i vjerojatnu biološku važnost) podskupa proteina sa 7 TM segmenata koji su mahom proteini iz obitelji fotosintetskih reakcijskih centara ili rodopsina.



Slika 48. Raspodjele broja lanaca (u postotcima) po podskupovima lanaca istog broja TM segmenata u početnim skupovima M1087, S481 i N1212.

Na slici 48 uočava se značajno veća zastupljenost lanaca s dva TM segmenta u skupu M1087 nego u ostala dva skupa. Odgovor leži u činjenici da je u tom skupu veliki broj lanaca potpuno identičnih slijedova prema bazi PDB. Kad se takvi slijedovi izuzmu, dobije se samo 430 lanaca (skup M430). Kada se pogledaju skupovi u kojima se nalaze samo proteinski slijedovi koji se razlikuju barem za jednu aminokiselinu, tada se dobije raspodjela prikazana na slici 49.

Vidi se da je u skupu N1212 broj lanaca u svim podskupinama znatno veći nego u drugim početnim skupovima. Naime, u posljednje dvije godineu bazu OPM [29] dodan je veliki broj novih lanaca s riješenom strukturom i položajima TM segmenata. Stoga, potrebno je kreirati novi reprezentativni skup polazeći od novijeg stanja baze OPM kako bi se provjerilo koliko je među novim riješenim strukturama takvih čiji aminokiselinski slijedovi imaju nisku međusobnu sličnost.



Slika 49. Raspodjele broja lanaca po podskupovima lanaca istog broja TM segmenata u početnim skupovima M430, S481 i N1212.

Kako bi se pokazala valjanost Algoritma 3 u tablici 33. prikazane su ukupne složenosti reprezentativnih skupova (dobivenih iz početnih skupova za prag identičnosti 20% ili sličnosti 30%), zajedno s reprezentativnim skupovima izabranim algoritmima (metodama) iz literature. U svim skupovima uočavaju se veće vrijednosti parametra složenosti skupa (entropijski koeficijent) za skupove izabrane Algoritmom 3. To je posebice vidljivo na primjeru reprezentativnih skupova izabranih iz skupa S392: (a) uz prag identičnosti 20%, gdje je ta razlika najveća (~20%), ili (b) uz prag sličnosti 30% gdje su dobivene vrijednosti više za oko 15%. Na skupovima M190 i M1087 rezultati su bolji prema složenosti za oko 1%. Na skupovima S148 i N189, rezultati su bolji za preko 10%. Najveća razlika je dobivena na skupu N263, gdje je naš izbor lanaca dao 28% veću kompleksnost skupa nego li izbor algoritmom UniqueProt [44].

Tablica 33. Usporedba ukupnih složenosti ($S_{0,uk}$) izabranih reprezentativnih skupova dobivenih Algoritmom 3 i algoritmima iz literature.

početni skup (prag identičnosti/sličnosti)	Algoritam 3		algoritam iz literature			omjer (%)	
	TM	$S_{0,uk}$		TM	$S_{0,uk}$	TM	$S_{0,uk}$
M190 (20% identičnosti)	509	2381	UniqueProt [42,44]	506	2367	100.59	100.59
S481 (20% identičnosti)	850	4284	Hobohm 2 [41]	820	4133	103.66	103.65
S481 (30% sličnosti)	621	3182	Hobohm 2 [41]	580	2979	107.07	106.81
S392 (20% identičnosti)	757	3806	Hobohm 2 [41]	638	3182	118.65	119.61
S392 (30% sličnosti)	560	2846	Hobohm 2 [41]	483	2462	115.94	115.60
M1087 (20% identičnosti)	546	2555	UniqueProt [43,44]	540	2529	101.11	101.03
S148 (20% identičnosti)	396	1913	UniqueProt [43,44]	352	1681	112.50	113.80
N189 (20% identičnosti)	426	2052	UniqueProt [43,44]	393	1853	108.40	110.74
N263 (20% identičnosti)	575	2717	UniqueProt [43,44]	463	2121	124.19	128.10
N907 (30% sličnosti)	942	4879					
N907 (20% identičnosti)	1514	7657					
N1212 (30% sličnosti)	1156	6113					
N1212 (20% identičnosti)	1846	9499					

Nadalje, ukupne složenosti reprezentativnih skupova dobivenih Algoritmom 3 za iste pragove identičnosti/sličnosti više su za preko 60%. Ovako velika razlika u ukupnoj složenosti skupova zbirni je doprinos same prirode početnog skupa (povezno datumom preuzimanja, jer kasniji datumi znače i veći broj proteina pohranjen u bazama), te kvalitete primijenjenog Algoritma 3.

Važnost razvoja novih i kvalitetnijih algoritama za izbor reprezentativnih skupova dobiva na značaju ako se uzme u obzir da je broj pronađenih novih vrsta strukture (engl. *fold-type*)u stagnaciji posljednjih godina. Stoga, i broj proteina u novim reprezentativnim skupovima raste sporo. Tako se, razvojem poboljšanih algoritama koji daju poboljšanja od samo 2 ili 3% u izboru reprezentativnog skupa, već dobiva značajno poboljšanje.

Rezultati dobiveni provedbom istraživanja u disertaciji značajno doprinose povećanju količine korisne (originalne) informacije o strukturama integralnih membranskih proteina α vrste niske međusobne sličnosti. Osjetno veća količina korisne informacije sadržana u većim i informativnijim reprezentativnim skupovima, temeljjeza razvoj boljih i pouzdanijih modela (računalnih metoda) za predviđanje strukture membranskih proteina α vrste. Također, veći reprezentativni skupovi membranskih proteina niske međusobne sličnosti korisni su za druge vrste analiza u kojima visoka međusobna sličnost zamagljuje ključnu i originalnu informaciju na temelju koje se žele donositi zaključci o njihovoj strukturi i funkciji. Dodatno, ukoliko se u analizama kreće od manjeg broja (reprezentativnih) proteinskih lanaca koji sadrže svu korisnu (originalnu) informaciju o strukturama nekog skupa proteina, skraćuje se vrijeme potrebno za provedbu analiza jer je potrebno proanalizirati (a ponekad i pregledati jednu po jednu) manji broj struktura proteinskih lanaca.

Saznanja do kojih se došlo radom na disertaciji imaju i općeniti značaj i primjenu, tj. mogu se primijeniti, s istom svrhom, na druge podskupine proteina. Stoga se očekuje kako će se i u tim slučajevima dobiti osjetno bolji rezultati koji mogu ubrzati i olakšati istraživački i stručni rad vezan uz analizu i donošenje zaključaka u istraživanjima u bio-znanostima.

4. ZAKLJUČAK

Provedena je analiza postojećih iz literature dostupnih algoritama za izbor reprezentativnog skupa integralnih membranskih proteina poznate strukture i utvrđeno kako je odlučivanje pri izboru pojedinih proteinskih lanaca uglavnom vezano za duljinu proteinskog lanca ili rezoluciju s kojom je riješena struktura. Svi ti kriteriji neovisni su o složenosti strukture lanca. Reprezentativni skupovi izabrani po navedenim kriterijima, sadrže veliki postotak lanaca jednostavne strukture sa samo jednim TM segmentom. Na temelju toga pretpostavljeno je da se definiranjem kriterija složenosti strukture i njihovom uporabom u izboru proteina u repezentativni skup može izabrati kvalitetniji skup (tj. skup veće složenosti, odnosno s više korisnih strukturnih informacija).

U početku istraživanja kao prvi kriterij nametnuo se broj TM segmenata u proteinskom lancu. Primjena tog kriterija u algoritmima za izbor repezentativnih skupova membranskih proteina ubrzo je dovela do izbora većih skupova proteinskih lanaca složenije strukture.

Nakon početne analize, uveden je koncept uravnoteženog nasumičnog modela za dva stanja sekundarne strukture,te je analizirana njegova primjena u procjeni stvarne točnosti metoda za predviđanje strukture membranskih proteina. Definiran je parametar za izračun najvjerojatnije točnosti uravnoteženog nasumičnog modela $Q_{2,rnd}$, koji se temelji na udjelu dviju vrsta sekundarnih struktura u proteinskom lancu. Vrijednost parametra $Q_{2,rnd}$ najmanja je kada je udio dviju vrsta sekundarne strukture u proteinu podjednak (oko 50%). Kada neki duži proteinski lanac sadrži samo jedan TM segment, udio jedne (pravilne) sekundarne strukture puno ja manji od 50%. U tom slučaju raste vrijednost najvjerojatnije nasumične točnosti $Q_{2,rnd}$. To znači da se za takvu strukturu već i nasumičnim modelom dobiva visoka točnost, što upućuje na to da je takva struktura jednostavnija za predviđanje. Međutim, parametar $Q_{2,rnd}$ dobro razlikuje lance prema složenosti strukture, ali samo za lance (približno) iste duljine.

Spomenuti nedostatak parametra točnosti $Q_{2,rnd}$ prevladan je uvođenjem koncepta segmentnog nasumičnog modela. Pritom, segment čini niz od d_i susjednih aminokiselina, a razmatrani su slučajevi kada između segmenata nema razmaka (r_{min} = 0), te kad postoje minimalni razmaci zadane duljine (broja aminokiselina) r_{min}.

U najjednostavnijem slučaju kada dopustimo duljinu segmenta od samo jedne aminokiseline, segmentni nasumični model zapravo postaje binomni nasumični model modelne strukture. Izraz za izračun broja mogućih realizacija (konformacija) modelne strukture u slučaju binomnog modela ekvivalentan jemodelu izjednačavanja koncentracija dvaju plinova iz statističke fizike. Nadalje, broj mogućih realizacija modelne strukture prema binomnom i općenito segmentnom modelu u vezi je s entropijom modelne strukture membranskog proteina, s tim što je slaganje bolje sa segmentnim modelom.

Izvedene su formule za izračunavanje broja mogućih realizacija modelne strukture membranskog proteina prema segmentnom modelu, i to za slučajeve kada:

a) je u modelnoj strukturi proizvoljni broj segmenata proizvoljne duljine,
b) u strukturi postoji više pod-skupina (klasa) segmenata podjednake duljine, pri čemu ne razlikujemo segmente unutar iste klase, te kad
c) između segmenata ne postoje razmaci, ili kad postoje razmaci proizvolje minimalne duljine r_{min}.

Broj mogućih realizacija modelne strukture doveden je u vezu sa složenošću sekundarne strukture membranskog proteina. Dodatno, prema statističkoj fizici, broj mogućih realizacija modelne strukture u vezi je s entropijom modelne strukture proteina, što je omogućilo i fizikalnu interpretaciju rezultata. Nadalje, vrijednosti tog parametra koji iskazuje složenost strukture povezane su (tj. značajno korelirane) s brojem TM segmenata u proteinskom lancu, kao i s duljinom lanca.

Parametri za iskazivanje složenosti strukture proteina korišteni su u razvoju poboljšanih algoritama za izbor proteinskih lanaca u reprezentativne skupove. Provedena je usporedba na više skupova membranskih proteina α vrste s algoritmima iz literatre (Hobohm 2 [41] i UniqueProt [44]). Razvijeni Algoritmi 1, 2 i 3 dali su veće i značajno kvalitetnije reprezentativne skupove membranskih proteina α vrste. Naposljetku, Algoritmi 1, 2 i 3 primijenjeni su u izboru novih reprezentativnih skupova membranskih proteina α vrste, polazeći od novih inačica baza OPM [29] i PDB [20] (srpanj 2017.). Taj je skup značajno veći(u nekim slučajevima i dvostruko) od svih skupova izdvojenih u posljednje četiri godine i objavljenih u literaturi (tablica 30). Jedan dio tog povećanja dolazi zbog:

a) povećanja baza proteinskih struktura u zadnje dvije godine, a drugi zbog
b) primjene kvalitetnijih algoritama razvijenih u disertaciji, koji se temelje na parametrima složenosti strukture membranskih proteina.

Među razvijenim algoritmima izdvaja se Algoritam 3 kvalitetom rezultata, brzinom rada i originalnošću izvedbe. Dodatno, taj algoritam nema analogije niti u jednom opisanom algoritmu u znanstvenoj literaturi.

Izabrani novi reprezentativni skupovi membranskih proteina α vrste niske međusobne sličnosti otvaraju brojne mogućnosti primjene u području modeliranja strukture proteina, kako za provjeru kvalitete i unapređenje postojećih modela, tako i za razvoj novih pouzdanijih modela. Te mogućnosti otvaraju se stoga što su novi skupovi osjetno veći i sadrže proteine složenije strukture a time i niže razine točnosti nasumičnog modela $Q_{2,rnd}$ (koje su bliže 50%). Na taj način otvara se veći prostor za unapređenja algoritama za predviđanje strukture membranskih proteina α vrste.

Parametri za iskazivanje složenosti strukture membranskih proteina pokazuju značajnu korelaciju s brojem TM segmenata. To omogućuje uporabu vrlo jednostavnog kriterija broja TM segmenata za grubu procjenu složenosti strukture membranskih proteina. Novi uvedeni koncepti nasumičnog i segmentnog nasumičnog modela te novi originalni algoritmi otvaraju mogućnosti (uz odgovarajuće prilagodbe) za njihovu primjenu u drugim područjima strukturne bioinformatike i molekulske biofizike, u području dizajniranja novih lijekova ili u znanosti o okolišu.

Etička načela

Planirana istraživanja nisu uključivala provedbu laboratorijskih pokusa, pa niti pokusa na laboratorijskim životinjama. U provedbi istraživanja poštivana su etička načela u znanstveno-istraživačkom radu.

Ustanove na kojima su provedena istraživanja u sklopu izrade doktorskoga rada

Istraživanja u sklopu izrade doktorskoga rada provedena su na Institutu Ruđer Bošković, Bijenička c. 54, Zagreb, Hrvatska i na Fakultetu prirodoslovno-matematičkih i odgojnih znanosti Sveučilišta u Mostaru, Matice hrvatske b.b., Mostar, Bosna i Hercegovina.

Znanstvena istraživanja provedena su uz financijsku potporu:

(1) Ministarstva znanosti i obrazovanja Republike Hrvatske u sklopu znanstvenog projekta br. 098-1770495-2919 (naziv projekta: Istraživanje odnosa između strukture i svojstava bioaktivnih molekula i proteina), i
(2) Zaklade HAZU (Hrvatska akademija znanosti i umjetnosti), Zagreb, Hrvatska.

Trošak školarine doktorskog studija podmirilo je Ministarstvo znanosti i obrazovanja Republike Hrvatske.

5. LITERATURA

[1] A. Krogh et al. *J. Mol. Biol.*, 305(3): 567-580, 2001.

[2] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore and J. Darnell. *Molecular cell biology*, 4th edition, New York: W. H. Freeman; 2000.

[3] J. Deisenhofer et al. *Nature*, 318: 618-624, 1985.

[4] P. Y. Chou and G. D. Fasman. *Biochemistry*, 13(2): 211-222, 1974.

[5] P. Y. Chou and G. D. Fasman. *Biochemistry*, 13(2): 222-245, 1974.

[6] J. Kyte and R. F. Doolittle. *J. Mol. Biol.*, 157(1): 105-132, 1982.

[7] D. Eisenberg et al. *J. Mol. Biol.*, 179(1): 125-142, 1984.

[8] G. von Heijne. *EMBO J.*, 5(11): 3021, 1986.

[9] G. von Heijne. *J. Mol. Biol.*, 225(2): 487-494, 1992.

[10] G. D. Fasman, ed., *Prediction of protein structure and the principles of protein conformations*, 359-389, New York and London, Plenum Press, 1989.

[11] D. Juretić et al. *Biopolymers*, 33(2): 255-273, 1993.

[12] B. Lučić, N. Trinajstić and D. Juretić, *Recognition of Membrane Protein Structure from Amino Acid Sequence, in: From Chemical Topology to Three-Dimensional Geometry*, (A. T. Balaban, Ed.), Plenum Publishing Corporation, New York, 117-158, 1997.

[13] D. Juretić et al. *Comput. Chem.*, 22(4): 279-294, 1998.

[14] D. Juretić, D. Zucić, B. Lučić and N. Trinajstić, *Protein Transmembrane Structure: Recognition and Prediction by Using Hydrophobicity Scales through Preference Functions, in Theoretical and Computational Chemistry, Theoretical Organic Chemistry*, (C. Parkanyi and W.C. Herndon Eds.), Amsterdam, Elsevier Science B. V., Volume 5., 405-445, 1998.

[15] D. Juretić, L. Zoranić and D. Zucić. *J. Chem. Inf. Comput. Sci.*, 42: 620-632, 2002.

[16] B. Rost and C. Sander. *J. Mol. Biol.*, 232(2): 584–599, 1993.

[17] B. Rost et al. *Protein Sci.*, 4(3): 521–533, 1995.

[18] J. A. Cuff et al. *Bioinformatics*, 14: 892–893, 1998.

[19] C. Cole, J. D. Barber and G. J. Barton. *Nucleic Acids Res.*, 36: W197–W201, 2008.

[20] H. M. Berman et al. *Nucleic Acids Res.*, 28: 235-242, 2000. (baza dostupna na adresi http://www.rcsb.org/pdb/home/home.do; poveznica za izdvajanje skupa reprezentativnih lanaca na (http://www.rcsb.org/pdb/statistics/clusterStatistics.do)

[21] C. Sander and R. Schneider. *Proteins*, 9(1): 56-68, 1991.

[22] S. F. Altschul et al. *Nucleic Acids Res.*, 25: 3389–3402, 1997.

[23] S. F. Altschul et al. *J. Mol. Biol.*, 215: 403–410, 1990.

[24] U. Hobohm et al. *Protein Sci.*, 1(3): 409–417, 1992.

[25] U. Hobohm and C. Sander. *Protein Sci.*, 3(3): 522–424, 1994.

[26] I. Sillitoe et al. *Nucleic Acids Res.*, 43(D1): D376-D381, 2014. (baza dostupna na adresi http://www.cathdb.info/, pristupljeno u rujnu 2017.)

[27] A. G. Murzin et al. *J. Mol. Biol.*, 247(4): 536-540, 1995. (baza dostupna na adresi http://scop.mrc-lmb.cam.ac.uk/scop/, a analiza strukturnih oblika dostupna na adresi http://www.proteinstructures.com/Structure/Structure/protein-fold.html, pristupljeno u rujnu 2017.)

[28] M. A. Lomize et al. *Nucleic Acids Res.*, 40(D1): D370-D376, 2012.

[29] M. A. Lomize et al. *Bioinformatics*, 22(5), 623-625, 2006. (baza dostupna na adresi opm.phar.umich.edu)

[30] A. Bairoch and B. Boeckmann. *Nucleic Acids Res.*, 19: 2247-2248, 1991.

[31] A. Bairoch and R. Apweiler. *Nucleic Acids Res.*, 28: 45-48, 2000.

[32] The UniProt Consortium. *Nucleic Acids Res.*, 43: D204-D212, 2015.

[33] G. E. Tusnady and I. Simon. *Bioinformatics*, 17(9), 849-850, 2001.

[34] C. P. Chen, A. Kernytsky and B. Rost. *Protein Sci.*, 11: 2774–2791, 2002.

[35] H. Zhou and Y. Zhou. *Protein Sci.*, 12(7): 1547–1555, 2003.

[36] H. Viklund and A. Elofsson. *Protein Sci.*, 13(7): 1908–1917, 2004.

[37] H. Viklund and A. Elofsson. *Bioinformatics*, 24: 1662-1668, 2008.

[38] A. Bernsel et al. *Nucleic Acids Res.*, 37: W465-W468, 2009.

[39] K. D. Tsirigos et al. *Nucleic Acids Res.*, 43: W401-W407, 2015.

[40] J. Batista and B. Lučić. *Proceedings of the First Adriatic Symposium on Biophysical Approaches in Biomedical Studies*, M. Raguž, K. Balaraman, T. Sarna, N. Ilić, D. Nejašmić, Danijel; J. Thelaner (Eds), Split, MedILS, 2014. 60-60.

[41] E. M. Rath et al. BMC *Bioinformatics*, 14(1): 111, 2013.

[42] J. Reeb et al. *Proteins*, 83: 473–484, 2015.

[43] M. Bernhofer et al. *Proteins*, 84: 1706–1716, 2016.

[44] S. Mika and B. Rost. *Nucleic Acids Res.*, 31(13): 3789–3791, 2003.

[45] J. Batista and B. Lučić, *Math/Chem/Comp 2016, 28th MC2 Conference, Book of Abstracts* / Vančik, Hrvoj and Cioslowski, Jerzy (ur.), Dubrovnik, 13-13, 2016. (poster).

[46] J. Batista, D. Vikić-Topić and B. Lučić. *Croat. Chem. Acta.*, 89(4): 527-534, 2016.

[47] Python 2 (dostupno na aresi: https://www.python.org/).

[48] M. F. Sanner. *J. Mol. Graph. Model.*, 17(1): 57-61, 1999.

[49] P. Rice, I. Longden and A. Bleasby. *Trends Genet.*, 16: 276-277, 2000. (program dostupan na adresi http://emboss.sourceforge.net/ ili na http://www.ebi.ac.uk/Tools/emboss/)

[50] H. M. Berman, K. Henrick and H. Nakamura. *Nat. Struct. Biol.*, 10(12): 980, 2003. (baza dostupna na adresi www.wwpdb.org)

[51] R. Leinonen et al. *Bioinformatics*, 20: 3236-3237, 2009. (baza dostupna na adresi http://www.uniprot.org/uniparc/)

[52] A. L. Lomize et al. *Protein Sci.*, 15: 1318-1333, 2006.

[53] A. L. Lomize, I. D. Pogozheva and H. I. Mosberg. *J. Chem. Inf. Model.*, 51: 930-946, 2011.

[54] G. E. Tusnády, Z. Dosztányi and I. Simon. *Bioinformatics*, 20: 2964–2972, 2004.

[55] G. E. Tusnády, Z. Dosztányi and I. Simon. *Nucleic Acids Res.*, 33: D275–D278, 2005.

[56] D. Kozma, I. Simon and G. E. Tusnády. *Nucleic Acids Res.*, 41: D524-D529, 2013. (baza dostupna na adresi http://pdbtm.enzim.hu/)

[57] G. E. Tusnady, Z. Dosztányi and I. Simon. *Bioinformatics*, 21: 1276–1277, 2005.

[58] M. Goujon et al. *Nucleic Acids Res.*, 38: W695-W699, 2010. (baza dostupna na adresi http://www.ebi.ac.uk)

[59] L. A. Bultet et al. *Nucleic Acids Res.*, 44(EPFL-ARTICLE-217833), D27-D37, 2016. (baza dostupna na adresi https://www.sib.swiss/)

[60] C. H. Wu et al. *Nucleic Acids Res.*, 31(1): 345-347, 2003. (baza dostupna na adresi http://pir.georgetown.edu/)

[61] B. E. Suzek et al. *Bioinformatics*, 23(10): 1282-1288, 2007.

[62] W. Kabsch and C. Sander. *Biopolymers*, 22: 2577–2637, 1983.

[63] S. B. Needleman and C. D. Wunsch. *J. Mol. Biol.*, 48(3): 443-453, 1970.

[64] B. Rost. *Protein Eng.*, 12: 85–94, 1999.

[65] B. Rasulev et al. *ACS Appl. Mater. Interfaces*, 9(2): 1781-1792, 2017.

[66] P. Baldi et al. *Bioinformatics*, 16(5): 412-424, 2000.

[67] J. G. Topliss and R. J. Costello. *J. Med. Chem.*, 15(10): 1066-1068, 1972.

[68] J. G. Topliss and R. P. Edwards. *J. Med. Chem.*, 22(10): 1238-1244, 1979.

[69] D. M. W. Powers. *J. Machine Learning Techn.*, 2(1): 37-63, 2011.

[70] I. Supek, *Teorijska fizika i struktura materije*, Zagreb, Školska knjiga, 1992.

[71] V. Šips, *Uvod u statističku fiziku*, Zagreb, Školska knjiga, 1990.

[72] D. K. Sunko, *Statistička fizika i termodinamika*, PMF Zagreb, 2016. (interna skripta)

6. ŽIVOTOPIS

Jadranko Batista rođen je 1977. godine u Zenici, BiH. Završio je srednju školu u Busovači 1995. godine i iste godine upisuje Pedagoški fakultet u Mostaru. Diplomirao je 2000. godine pod mentorstvom dr. sc. Željka Antunovića na temu "Magnetski monopoli" i stječe zvanje profesor matematike i fizike. Godine 2008. upisuje doktorski studij biofizike na PMF-u Sveučilišta u Splitu.

Trenutačno radi na Fakultetu prirodoslovno-matematičkih i odgojnih znajosti Sveučilišta u Mostaru kao viši asistent na Studiju fizike gdje sudjeluje u nastavi na kolegijima: Opća fizika 4, Astronomija i astofizika, Klasična mehanika, Kvantna fizika, te na drugim studijima: Uvod u opću fiziku, Fizika, Fizika 1 i Fizika 2.

Kao autor ili koautor objavio je 3 rada od kojih je jedan rad u časopisima indeksiranim u bazi Current Contents (kao prvi autor), a u obliku postera izvještavao je na desetak konferencija. Kao suautor objavio je šest udžbenika/priručnika fizike za osnovnu školu, i jedan udžbenik za studente fizike „Mehanika: metodička zbirka zadataka s rješenjima“.

7. POPIS PUBLIKACIJA

Izvorni znanstveni i pregledni radovi u CC časopisima

1. Batista, Jadranko; Vikić-Topić, Dražen; Lučić, Bono.
The difference between the accuracy of real and the corresponding random model is a useful parameter for validation of two-state classification model quality. // Croatica chemica acta. 89 (2016), 4; 527-534 (članak, znanstveni).

Znanstveni radovi u drugim časopisima

1. Lučić, Bono; Sović, Ivan; Batista, Jadranko; Skala, Karolj; Plavšić, Dejan; Vikić-Topić, Dražen; Bešlo, Drago; Nikolić, Sonja; Trinajstić, Nenad.
The Sum-Connectivity Index - An Additive Variant of the Randić Connectivity Index. // Current computer-aided drug design. 9 (2013), 2; 184-194 (članak, znanstveni).

Radovi u zbornicima skupova

1. Glunčić, M.; Paar, V.; Basar, I.; Vlahović, I.; Rosandić, M.; Dekanić, K.; Citković, M.; Jelovina D.; Paar, P.; Kelić, A.; Batista, J.
Direct mapping of symbolic DNA sequence into frequency domain and identification of higher order repeats // Bioinformatics and biological physics: proceedings of the scientific meeting / Vladimir Paar (ur.) / Zagreb: Hrvatska akademija znanosti i umjetnosti, 2013. 17-46 (predavanje, domaća recenzija, objavljeni rad).

Sažeci u zbornicima skupova

1. Batista, Jadranko; Lučić, Bono.
Quanitification of complexity of integral membrane protein secondary structure // Proceedings of the Second Adriatic Symposium on Biophysical Approaches in Biomedical Studies / Raguž, Marija; Kalyanaramam, Balaraman; Sarna, Tadeusz; Ilić, Nada; Nejašmić, Danijel; Thelaner, Jane (ur.) / Split, Hrvatska: Mediterranean Institute for Life Sciences, 2017. 80-80 (poster, međunarodna recenzija, sažetak, znanstveni).

2. Lučić, Bono; Batista, Jadranko; Papeš-Šokčević, Lidija; Nadramija, Damir.
The outcome of reasoning based on models greatly depends on the procedure used for their validation // Math/Chem/Comp 2017, Book Of Abstracts / Vančik, Hrvoj; Cioslowski, Jerzy (ur.) / Dubrovnik; Hrvatska, 2017. 28-28 (predavanje, međunarodna recenzija, sažetak, znanstveni).

3. Batista, Jadranko; Lučić, Bono.
Estimation of chance accuracy in classification structure-property models // Math/Chem/Comp 2016, 28th MC2 Conference, Book of Abstracts / Vančik, Hrvoj; Cioslowski, Jerzy (ur.) / Dubrovnik, 2016. 13-13 (poster, međunarodna recenzija, sažetak, znanstveni).

4. Batista, Jadranko; Lučić, Bono.
Influence of Differences in Experimental Structure Annotations on Accuracy of Structure Prediction of Membrane Proteins // Proceedings of the First Adriatic Symposium on Biophysical Approaches in Biomedical Studies / Raguž, Marija; Kalyanaramam, Balaraman; Sarna, Tadeusz; Ilić, Nada; Nejašmić, Danijel; Thelaner, Jane (ur.) / Split: Mediterranean Institute for Life Sciences, 2014. 60-60 (poster, međunarodna recenzija, sažetak, znanstveni).

5. Brana, Josip H.; Batista, Jadranko.
Kanonski formalizam za temeljna polja i spin-spin međudjelovanja // Knjiga sažetaka 8. znanstvenog sastanka Hrvatskog fizikalnog društva / M. Požek i dr. (ur.) / Primošten: Hrvatsko fizikalno društvo, 2013. 130-130 (poster, sažetak, znanstveni).

6. Batista, Jadranko; Lučić, Bono.
Significance verification and simplification of some often used structure-property models in molecular biosciences // Book of Abstracts, The 3rd Adriatic Meeting on Computational Solutions in Life Sciences / Babić, Darko; Došlić, Nadja; Smith, David; Tomić, Sanja; Vlahoviček, Kristian (ur.) / Zagreb: Centre for Computational Solutions in the Life Sciences, 2009. (poster, sažetak, znanstveni).

7. Lučić, Bono; Batista, Jadranko; Vikić-Topić Dražen; Plavšić, Dejan.
Procjena razine slučajne korelacije pri modeliranju strukturnih svojstava proteina // Knjiga sažetaka 6. znanstvenog sastanka Hrvatskog fizikalnog društva / Buljan, Hrvoje; Horvatić, Davor (ur.) / Primošten: Hrvatsko fizikalno društvo, 2009. 174-174 (poster, sažetak, znanstveni).

8. Lučić, Bono; Batista, Jadranko; Juretić, Davor.
Poboljšanje modela za predviđanje kristalizacije proteina // Knjiga sažetaka 5. znanstvenog sastanka Hrvatskog fizikalnog društva / A. Dulčić i dr. (ur.) / Zagreb: Fizički odsjek, PMF Zagreb, 2007. 106-106 (poster, sažetak, znanstveni).

9. Lučić, Bono; Batista, Jadranko.
Predviđanje udjela sekundarne strukture u proteinima // Četvrti znanstveni sastanak Hrvatskog fizikalnog društva, Knjiga sažetaka / Kumerički, Krešimir et al. (ur.) / Zagreb: PMF, 2003. 163-163 (poster, sažetak, znanstveni).

Znanstveni radovi u pripremi

1. Batista, Jadranko and Lučić, Bono, *Minimal, maximal and the most probable classification accuracy of a two-state random model*, rad u pripremi

2. Batista, Jadranko and Lučić, Bono, *Improved selection of non-redundant representative set of integral membrane protein chains of alpha-type by using protein structural attributes*, rad u pripremi

3. Batista, Jadranko and Lučić, Bono, *Segmental model estimates the number of realizations of model structure of membrane proteins*, rad u pripremi

8. PONOVLJENI OSNOVNI PODACI BEZ POTPISA

Sveučilište u Splitu, Prirodoslovno-matematički fakultet

Odjel za fiziku, Poslijediplomski sveučilišni doktorski studij Biofizika

"Izbor reprezentativnog skupa membranskih proteina poznate strukture: razvoj poboljšanih algoritama uporabom koncepta nasumičnog modela"

Doktorski rad autora Jadranka Batiste kao dio obaveza potrebnih za stjecanje doktorata znanosti, izrađen pod vodstvom mentora dr. sc. Bone Lučića, višeg znanstvenog suradnika.

Dobiveni akademski naziv i stupanj: doktor iz područja prirodnih znanosti, polje fizika.

Povjerenstvo za ocjenu doktorskog rada u sastavu:

1. prof. dr. sc. Mile Dželalija, red. prof.
2. prof. dr. sc. Paško Županović, red. prof.
3. dr. sc. Sanja Tomić, znanstvena savjetnica
4. dr. sc. Igor Weber, znanstveni savjetnik, zamjenski član

1. _____

2. _____

3. _____

Potvrđuje da je disertacija obranjena dana 05. 01. 2018.

Voditelj studija: prof. dr. sc. Paško Županović

_____

Predsjednik vijeća studija: prof. dr. dr. h. c. Vlasta Bonačić-Koutecky

_____

9. SAŽETAK

Strukturu membranskih proteina osjetno je teže eksperimentalno odrediti nego strukturu topljivih proteina. Zbog toga je u proteinskoj bazi podataka pohranjeno daleko manje riješenih struktura membranskih proteina u odnosu na njihovu procijenjenu zastupljenost u genomima živih organizama. Taj se nesrazmjer nastoji smanjiti razvojem računalnih metoda za predviđanje strukture i topologije membranskih proteina. Kako bi se razvio pouzdani model za predviđanje strukture membranskih proteina, potrebno je provesti njegovu optimizaciju na što većem (reprezentativnom) skupu proteina poznatih struktura, među kojima je najviša dopuštena sličnost u primarnim strukturama ispod 30%. Postojeći algoritmi nastoje izabrati u reprezentativne skupove integralnih membranskih proteina alfa vrste uglavnom proteine čija je struktura riješena s boljom rezolucijom. Uočava se da postojeći algoritmi u reprezentativne skupove učestalije izabiru proteine s malim brojem transmembranskih segmenata. Pritom se u radu algoritama ne razmatraju kao bitni čimbenici i pokazatelji složenosti strukture proteina. S druge strane, za očekivati je da će modeli biti pouzdaniji u predviđanju ako su razvijeni na skupu proteina složenijih struktura.

S ciljem kvantificiranja složenosti strukture uveden je koncept nasumičnog modela s dvije sekundarne strukture (s dva stanja). Pritom, jedno stanje odgovara dijelu lanca koji je u membrani u pravilnoj strukturi uzvojnice α, a drugo odgovara nepravilnoj strukturi aminokiselina u ostatku lanca. Nadalje, izveden je izraz za procjenu točnosti nasumičnog modela za dva stanja, za koju je uočeno da je u vezi sa složenošću strukture. Slaba strana tog modela nemogućnost je preciznog razlikovanja složenosti struktura lanaca koji se značajnije razlikuju po duljini. Kako bi se definirali kriteriji složenosti strukture osjetljivi na duljinu, uvedeni su koncepti binomnog i segmentnog nasumičnog modela i izvedeni izrazi za izračun broja mogućih realizacija modelne sekundarne strukture proteina.

U binomnom nasumičnom modelu promatra se broj mogućih realizacija strukture proteina kada se aminokiseline pojedinačno mogu nalaziti u jednom od dva stanja (stanje u membrani, i stanje izvan membrane). Nedostatak tog nasumičnog modela u tome je što daje preveliku težinu duljim proteinskim lancima, bez obzira na omjer brojeva aminokiselina u ili izvan membrane.

Binomni nasumični model nadograđen je uzimajući u obzir činjenicu da su transmembranski segmenti u integralnim membranskim proteinima alfa vrste najčešće duljine 18-22 aminokiseline. Promatrajući i analizirajući nasumično razmještanje određenog broja segmenata sekundarne strukture u lancu, uočeno je da broj mogućih realizacija modelne strukture vjernije opisuje složenost strukture.

Definiran je segmentni nasumični model i izveden izraz za izračun broja mogućih nasumičnih realizacija modelne strukture membranskog proteina za slučaj kad između segmenata postoji ili kad ne postoji minimalni dopušteni razmak. Provedena je usporedba između razvijenih koncepata nasumičnih modela koja je pokazala da je broj mogućih realizacija modelne strukture prema segmentnom nasumičnom modelu u vezi sa složenošću strukture i odgovara fizikalnoj definiciji entropije. Nadalje, broj mogućih realizacija u značajnoj je korelaciji s brojem transmembranskih segmenata u proteinima.

Saznanja do kojih se došlo s pomoću uvedenih koncepata nasumičnih modela i iz usporedbi na više skupova membranskih proteina, ugrađeno je u razvijene algoritme za izbor reprezentativnih skupova membranskih proteina alfa vrste. Algoritmi su uspoređeni međusobno kao i s algoritmima iz literature na više skupova membranskih proteina poznate strukture. Pritom, najboljim se pokazao Algoritam 3 koji, u svakom koraku izvođenja, donosi odluku o tome koji se protein izbacuje a koji zadržava u reprezentativnom skupu na temelju originalne analize broja zajedničkih susjeda između proteinskih lanaca u početnom skupu. Razvijeni algoritmi izabiru u reprezentativni skup membranske proteine značajno složenije strukture s približno 5 - 20% većim ukupnim brojem transmembranskih segmenata. Primjene razvijenih algoritama na postojeće baze membranskih proteina poznate strukture daju reprezentativne skupove niske međusobne sličnosti koji su po broju lanaca, transmembranskih segmenata i entropijskom koeficijentu dobivenom segmentnim nasumičnim modelom veći za više od 50% od skupova iz literature. Ti su skupovi veći značajnijim dijelom zbog većih baza i većeg početnog skupa. Međutim, značajan doprinos kvaliteti reprezentativnog skupa dolazi zbog postupka izbora provednog poboljšanim algoritmima razvijenim u disertaciji. Strukture proteinskih lanaca u izabranim reprezentativnim skupovima značajno su složenije u odnosu na strukture membranskih proteina alfa vrste iz skupova objavljenih u literaturi.

10. ABSTRACT

It is more difficult to determine experimentally the structure of membrane protein than that of soluble protein. That is why the number of solved structures of membrane proteins in databases is much less than the estimated number of genes present in genomes of living organisms that are for coding membrane proteins. This disproportion is tended to be reduced by development of computational methods for prediction of structure and topology of membrane proteins. In order to develop a reliable model for predicting membrane protein structure, it is necessary to perform model optimisation on the largest (representative) set of membrane proteins of known structures with mutual similarities of their primary structures below 30%. Existing algorithms tend to select into representative sets of integral membrane proteins of alpha-type mostly protein chains whose structures are solved with a better resolution. Also, it is perceived that existing algorithms, in higher percentage, select into representative sets proteins having only few (one or two) transmembrane segments. Doing in such a way, complexity of structure is not considered (at all) as an important factor and indicator of protein structure complexity. On the other hand, it is for expected that models could be more reliable in prediction if they will be developed on a set of proteins having more complex structure.

In order to quantify structure complexity, a random model concept with two secondary structure (two states) was introduced. In that case, one state corresponds to regular α-helix structure of membrane part (segment), and the second state corresponds to irregular structure of amino acids in extra-membrane parts. Further, a formula for estimation of accuracy of two-state random model is derived, and perceived that it is in relation with the complexity of structure. A weakness of that model is faced in inability to distinguish properly complexities of structures of chains of considerably different lengths. I order to define improved criteria of structure complexity, the binomial and segmental random model concepts were introduced. Moreover, corresponding formulae for the calculation of the number of possible realizations of protein model secondary structure, showing the analogy with entropy, were devised.

In the binomial random model the number of possible realizations of protein structure in such a way that single amino acids are considered to be in one of two stucture states (in membrane or in extra-membrane part). The deficiency of this model is that it gives too large weight to larger proteins, regardless the ratio of the total numbers of amino acids present in membrane and in extra-membrane parts.

The binomial random model is upgraded by taking into account the fact that transmembrane segments in alpha-type integral membrane proteins are mostly of 18-22 amino acids in length. Considering and analysing random positioning of a certain number of secondary structure segments in a protein chain, it is perceived that the number of possible realizations of model structure is an adequate measure of structure complexity.

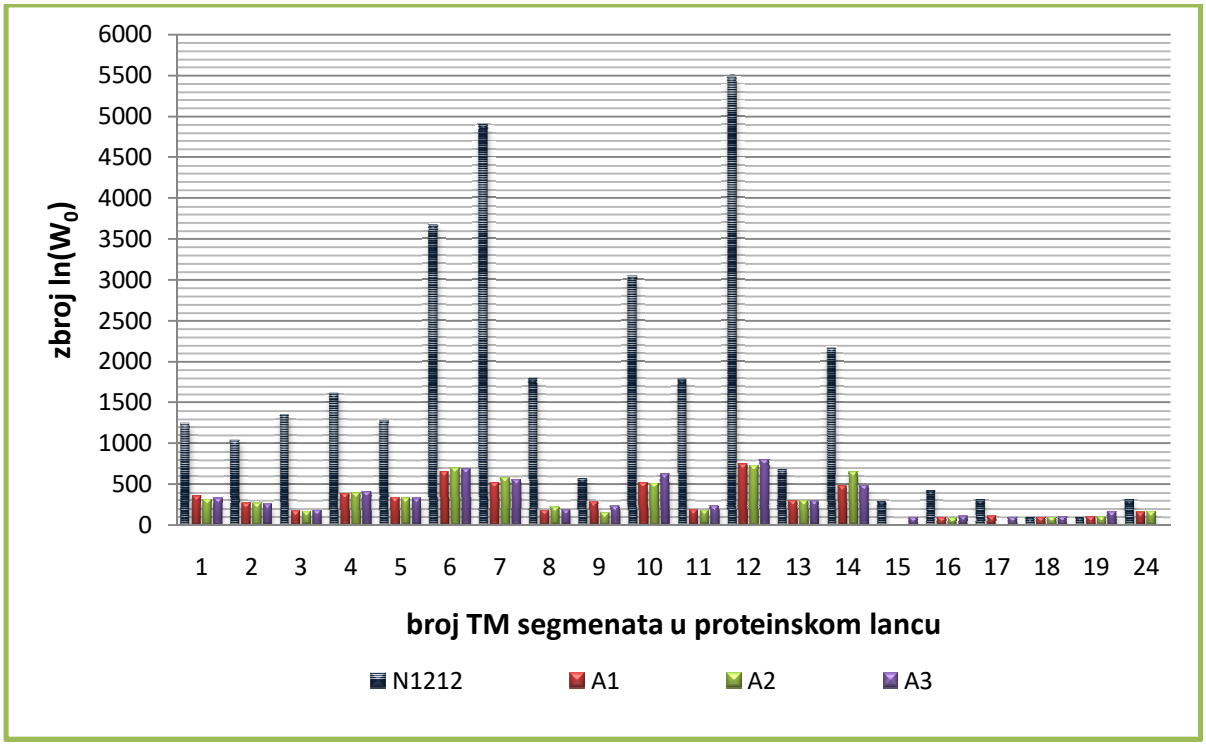
That is why a segmental random model is defined and a formula for the computation of the number of possible random realizations of model structure of membrane protein in case when a minimal space exists (or does not exist) between segments. A comparative analysis between developed random model concepts is performed showing that the number of possible realizations of model structure according to segmental random model is related to the structure complexity and corresponds to physical definition of entropy. In addition, the number of realizations is strongly correlated with the number of transmembrane segments in proteins.

These results obtained through the analysis of novel random model concepts and from the comparisons on several sets of membrane proteins are used in development of algorithms for selection of representative sets of alpha-type membrane proteins. Developed algorithms are compared both mutualy and with algorithms from literature on several sets of membrane proteins of known structure. In comparisons, the best overall performance shows the Algorithm 3 which, in each step, brings decision which of proteins will be eliminated and which one will be kept in representative set based on an original analysis of the number of common neighbours between protein chains from the initial set. Developed algorithms select in representative sets membrane proteins of significantly more complex structures, with the total number of transmembrane segments larger for ~ 5 - 20%.

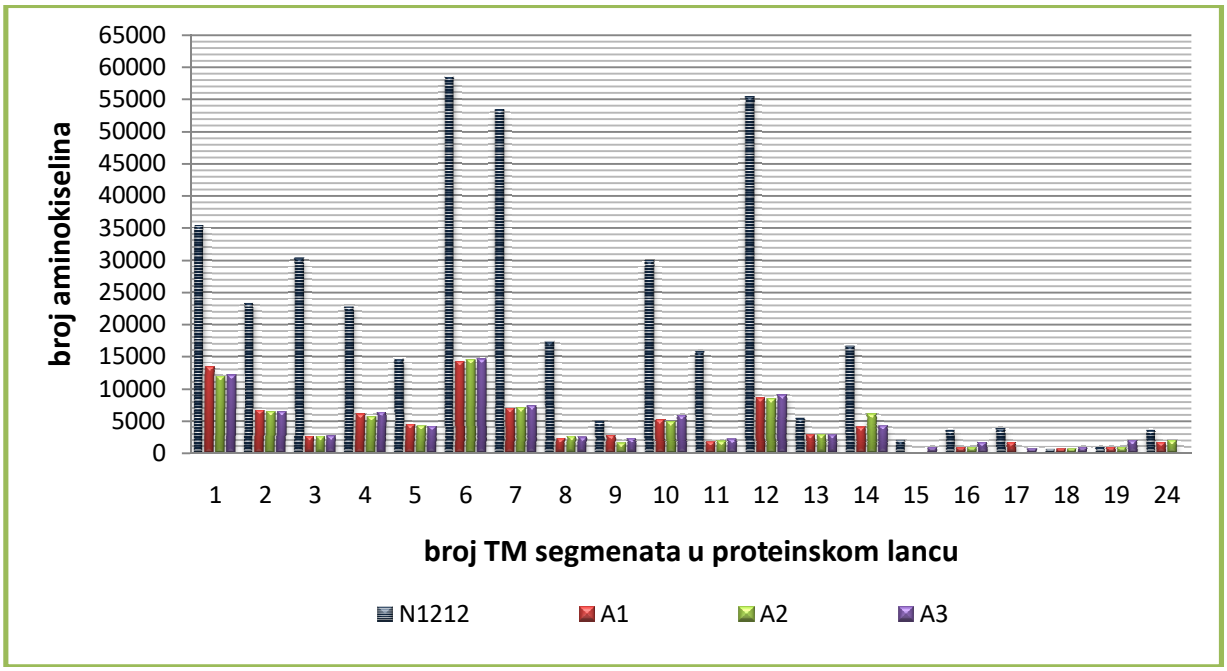
The applications of developed algorithms on existing membrane protein database of known structure give representative sets of low mutual similarity which are (in total) according to the number of chains, transmembrane segments and entropy coefficients obtained by the segmental random model significantly better for (in average) more than 50% than the results from literature. These data sets are larger because many proteins are deposited in protein databases in last two years. However, a significant contribution to the quality of representative set comes from improved algorithms developed in dissertation. Structures of protein chains in selected representative sets are significantly more complex comparing with the structures of alpha-type membrane proteins in published data sets from literature.

11. PRILOZI

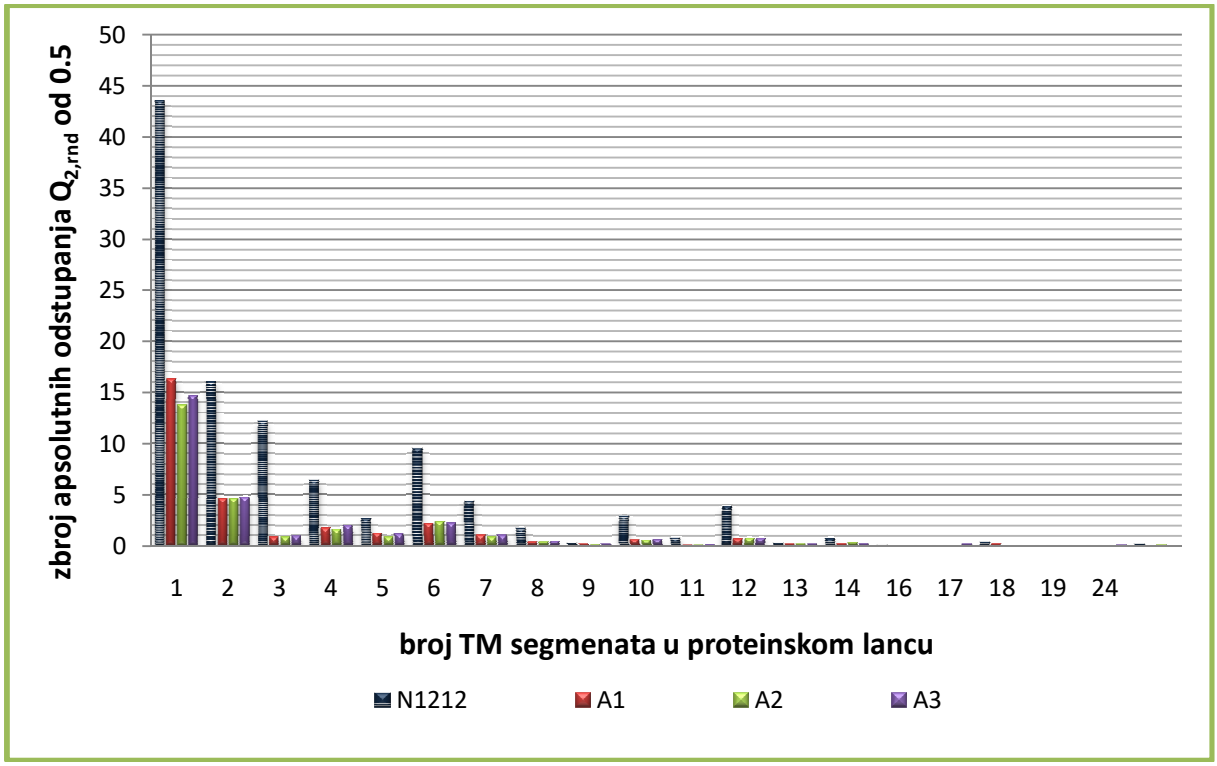
A. Slike uz analizu obilježja skupa membranskih proteina N1212 i izabranih reprezentativnih podskupova



Slika 1.A. Ukupni iznos entropijskog koeficijenta $S_{0,uk}$ u podskupovima lanaca istog broja TM segmenata za početni skup N1212 i skupove izabrane Algoritmima 1, 2 i 3 (za prag sličnosti 30%).

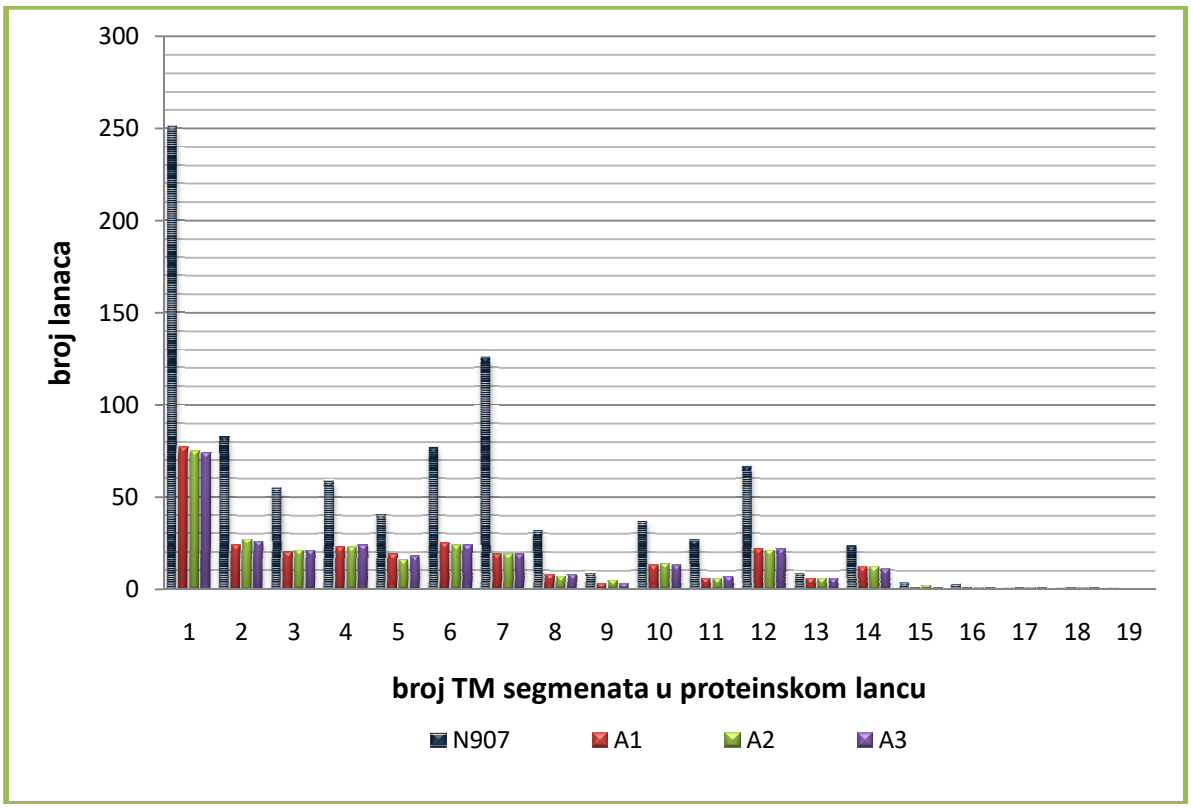


Slika 2.A. Ukupni broj aminokiselina u podskupovima lanaca istog broja TM segmenata za početni skup N1212 i skupove izabrane Algoritmima 1, 2 i 3 (za prag sličnosti 30%).

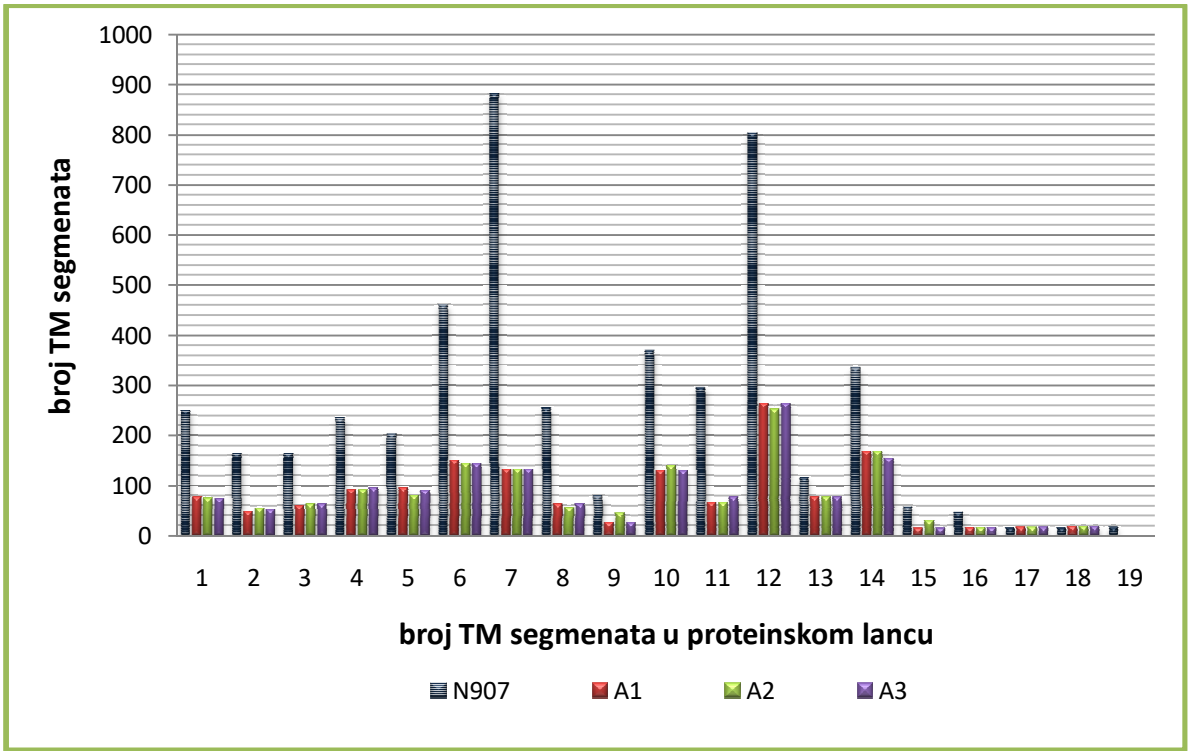


Slika 3.A. Ukupni iznos koeficijenta ($Q_{2,rnd} - 0.5$) u podskupovima lanaca istog broja TM segmenata za početni skup N1212 i skupove izabrane Algoritmima 1, 2 i 3 (za prag sličnosti 30%).

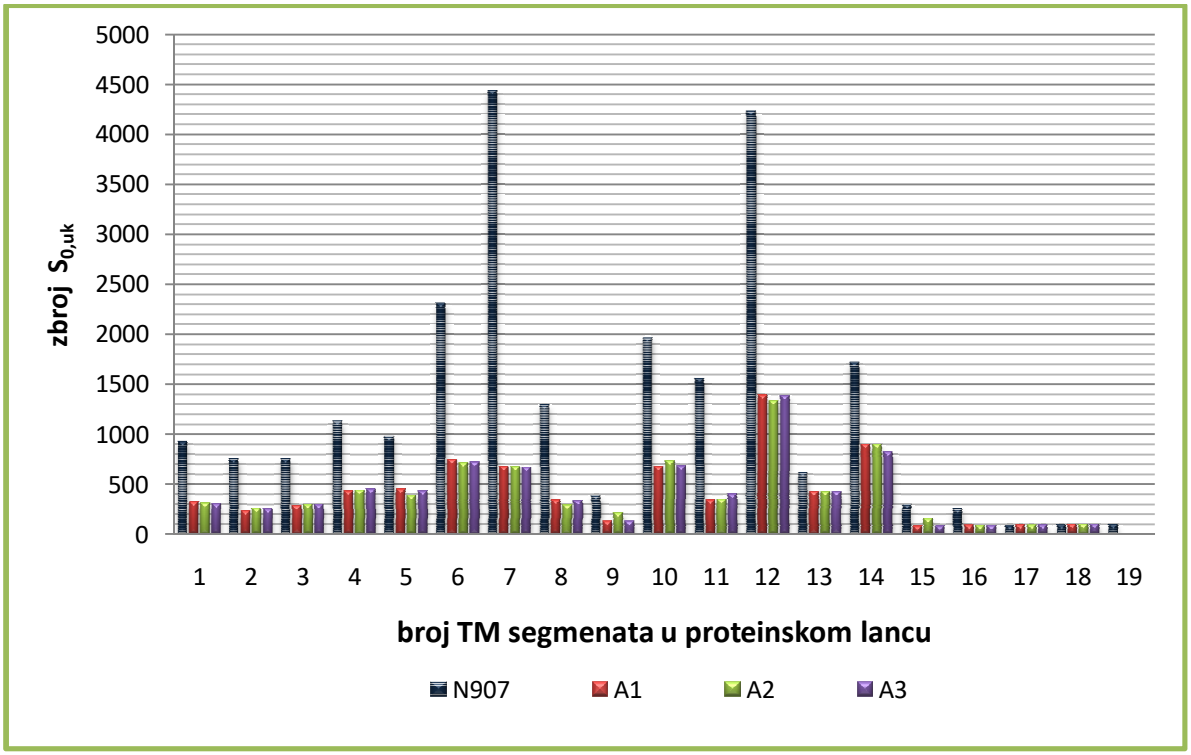
B. Slike uz analizu obilježja skupa membranskih proteina N907 i izabranih reprezentativnih podskupova



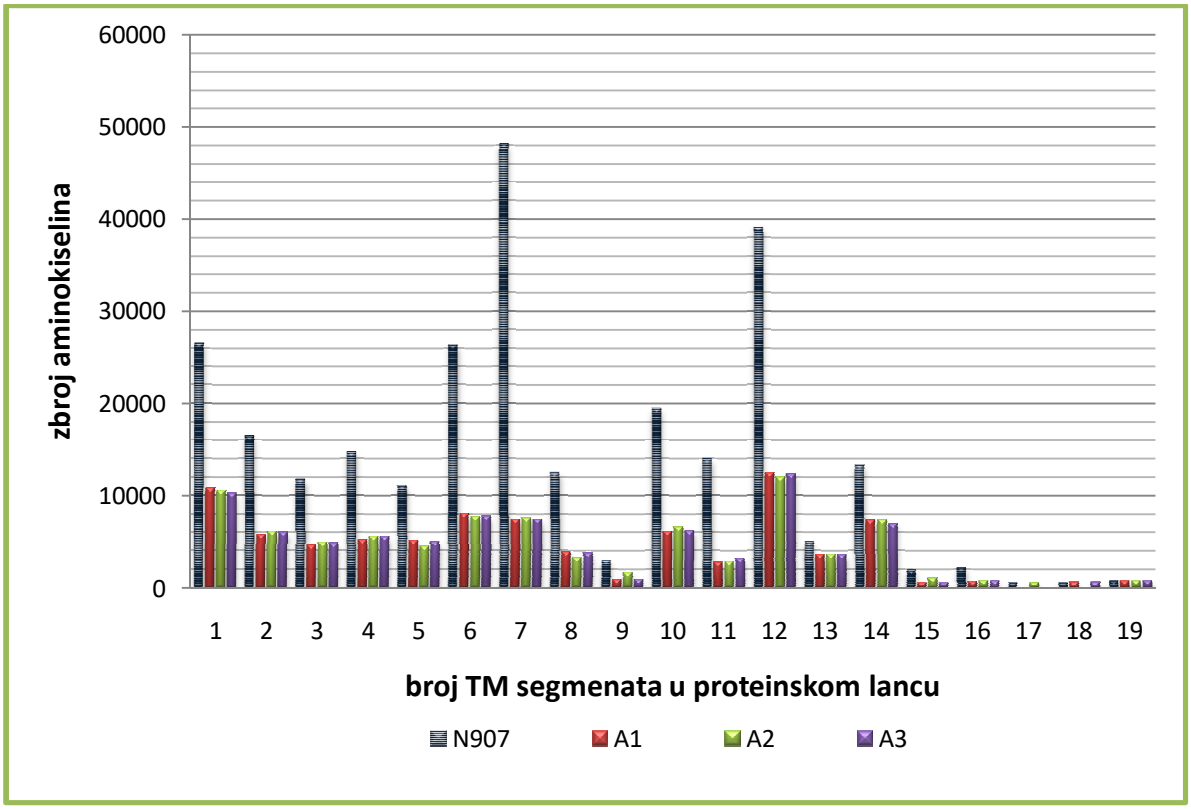
Slika 1.B. Broj lanaca u podskupovima lanaca istog broja TM segmenata za početni skup N907 i skupove izabrane Algoritmima 1, 2 i 3 (za prag identičnosti 20%).



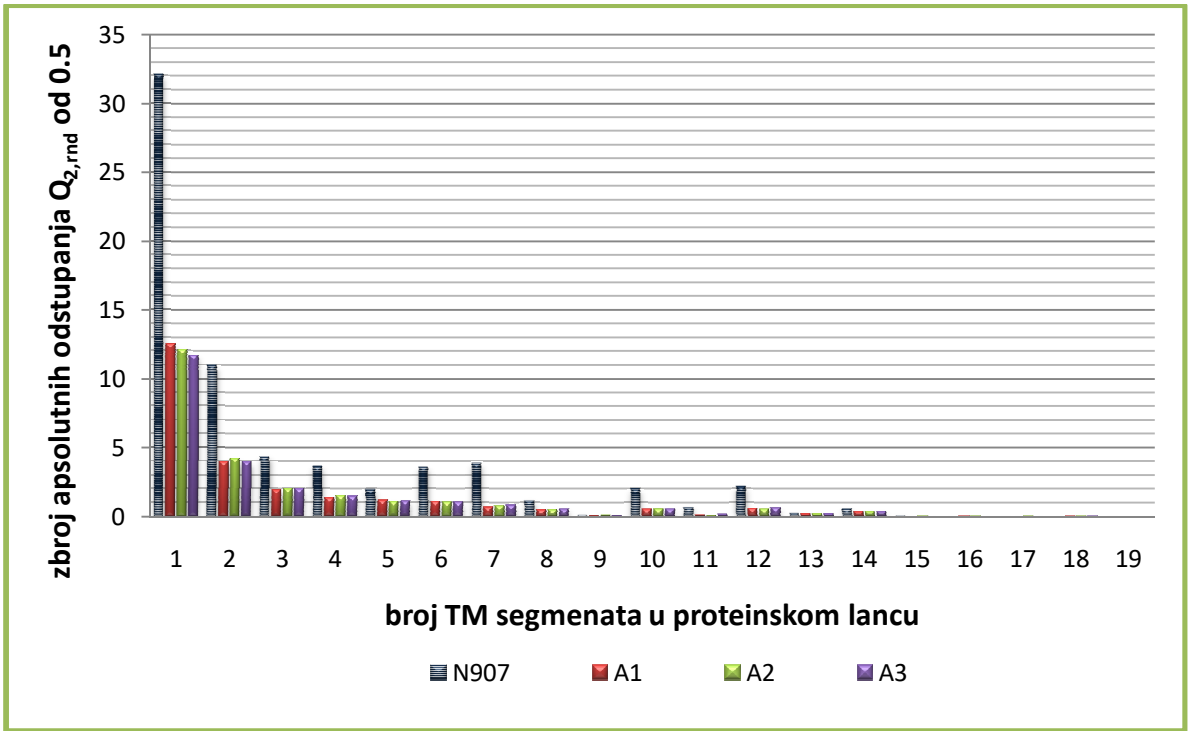
Slika 2.B. Ukupni broj TM segmenata u podskupovima lanaca istog broja TM segmenata za početni skup N907 i skupove izabrane Algoritmima 1, 2 i 3 (za prag identičnosti 20%).



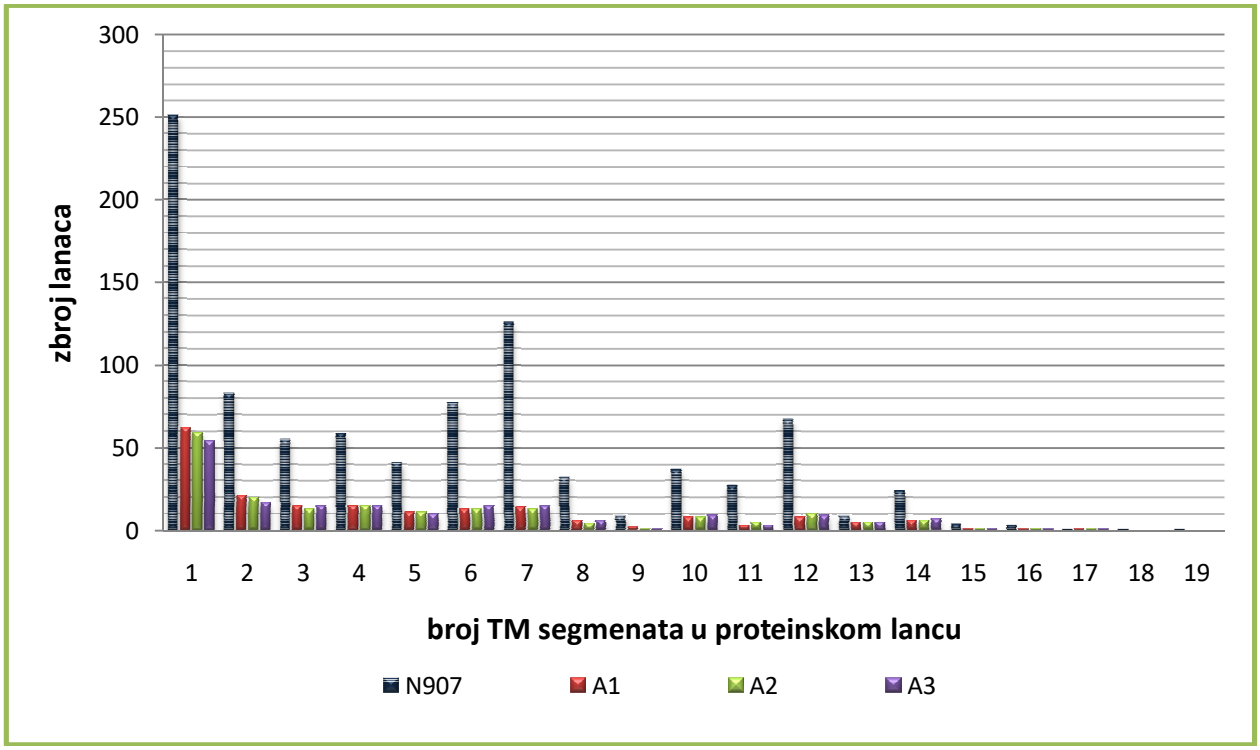
Slika 3.B. Ukupni iznos entropijskog koeficijenta $S_{0,uk}$ u podskupovima lanaca istog broja TM segmenata za početni skup N907 i skupove izabrane Algoritmima 1, 2 i 3 (za prag identičnosti 20%).



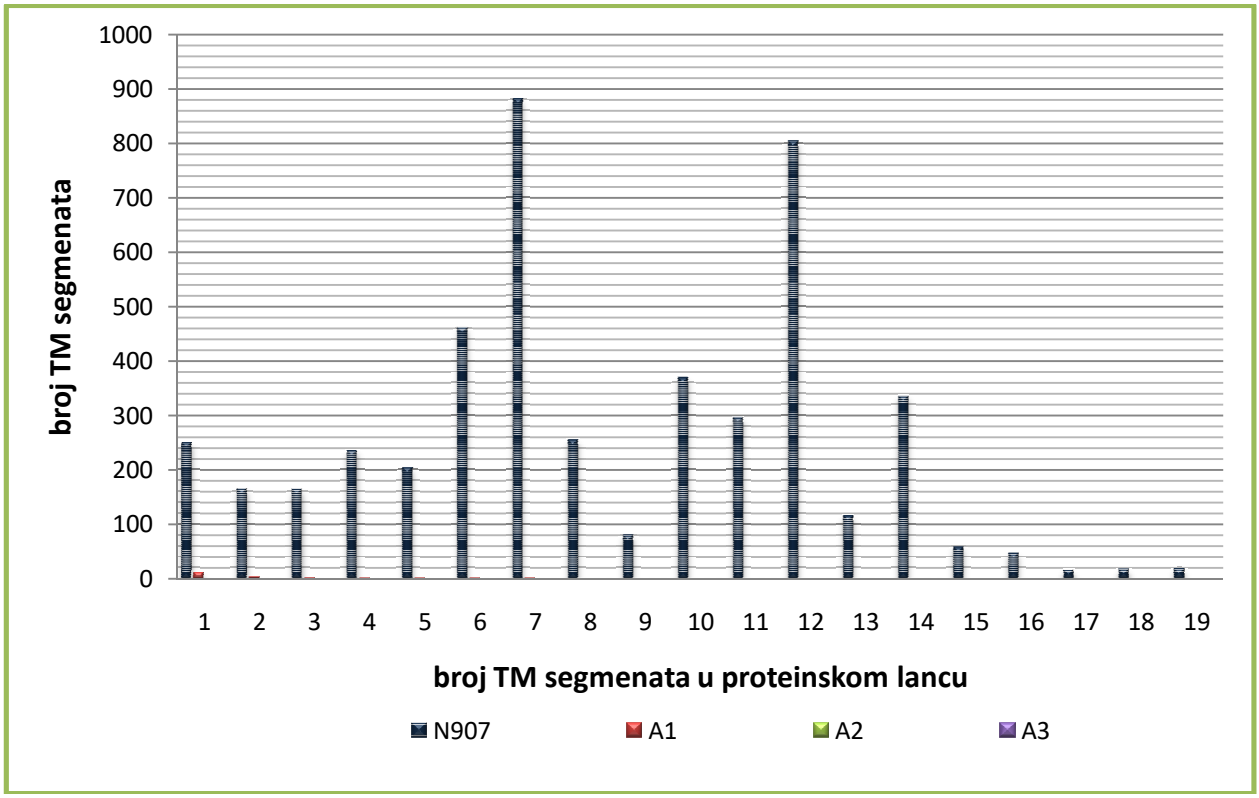
Slika 4.B. Ukupni broj aminokiselina u podskupovima lanaca istog broja TM segmenata za početni skup N907 i skupove izabrane Algoritmima 1, 2 i 3 (za prag identičnosti 20%).



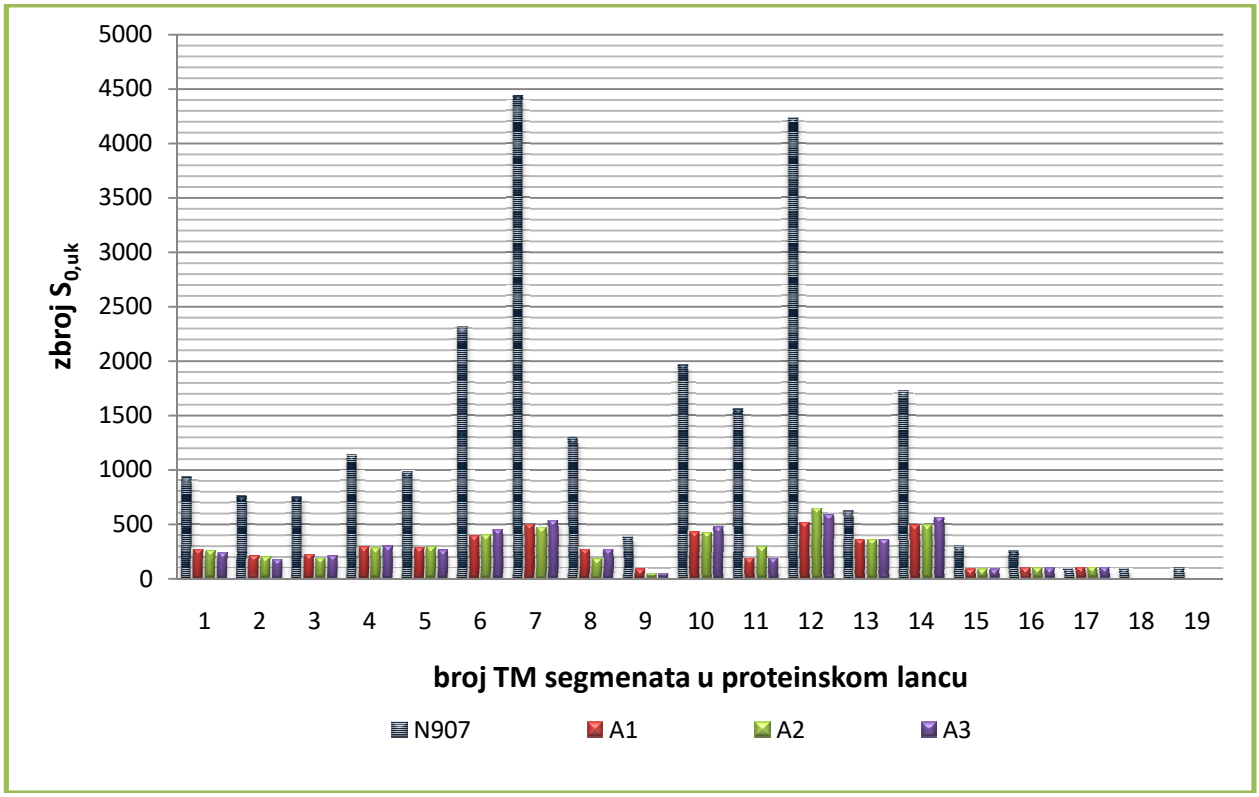
Slika 5.B. Ukupni iznos koeficijenta ($Q_{2,rnd} - 0.5$) u podskupovima lanaca istog broja TM segmenata za početni skup N907 i skupove izabrane Algoritmima 1, 2 i 3 (za prag identičnosti 20%).



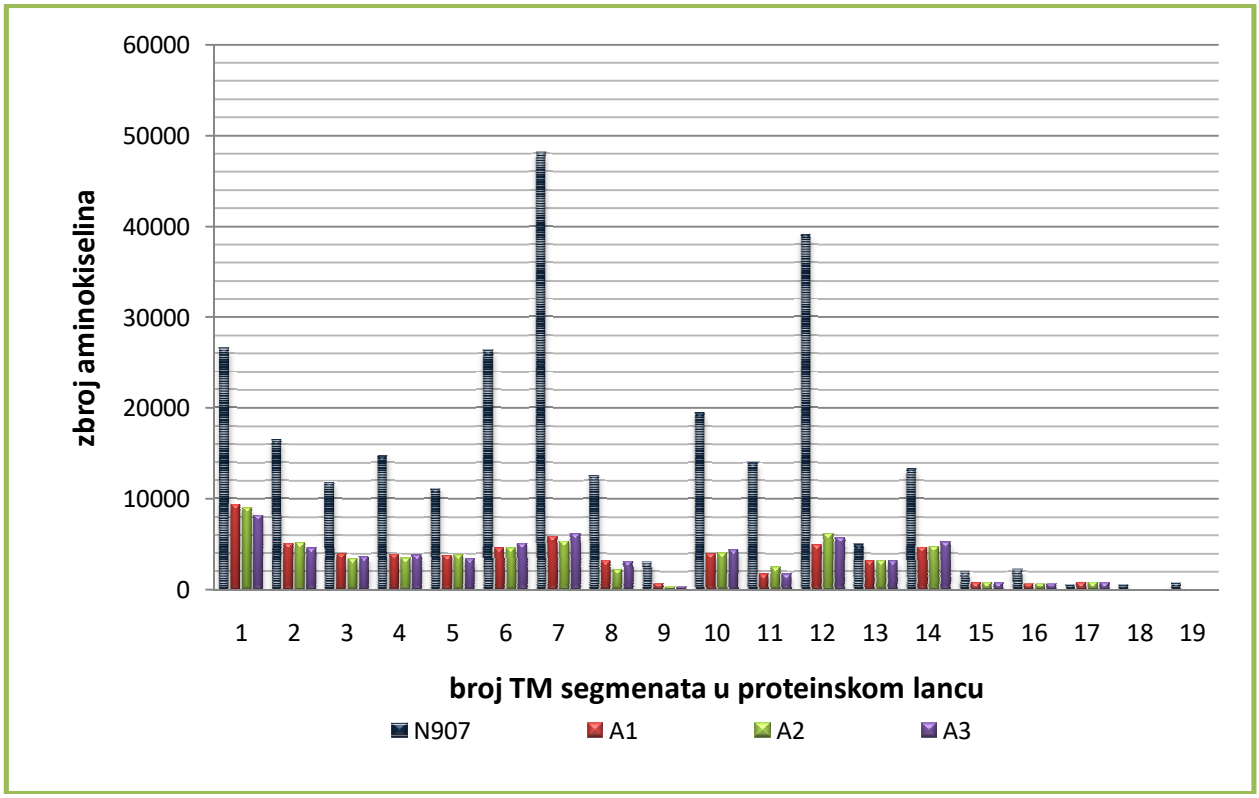
Slika 6.B. Broj lanaca u podskupovima lanaca istog broja TM segmenata za početni skup N907 i skupove izabrane Algoritmima 1, 2 i 3 (za prag sličnosti 30%).



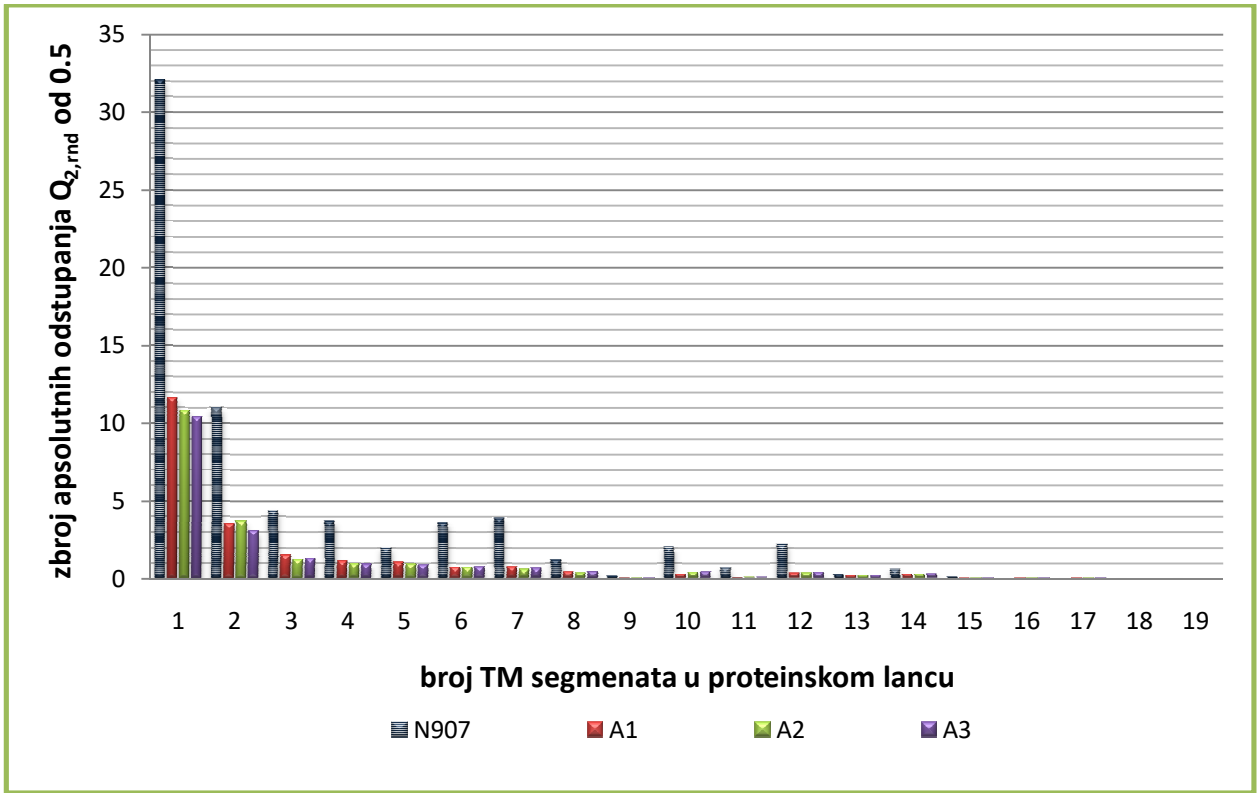
Slika 7.B. Ukupni broj TM segmenata u podskupovima lanaca istog broja TM segmenata za početni skup N907 i skupove izabrane Algoritmima 1, 2 i 3 (za prag sličnosti 30%).



Slika 8.B. Ukupni iznos entropijskog koeficijenta $S_{0,uk}$ u podskupovima lanaca istog broja TM segmenata za početni skup N907 i skupove izabrane Algoritmima 1, 2 i 3 (za prag sličnosti 30%).



Slika 9.B. Ukupni broj aminokiselina u podskupovima lanaca istog broja TM segmenata za početni skup N907 i skupove izabrane Algoritmima 1, 2 i 3 (za prag sličnosti 30%).



Slika 10.B. Ukupni iznos koeficijenta $(Q_{2,rnd} - 0.5)$ u podskupovima lanaca istog broja TM segmenata za početni skup N907 i skupove izabrane Algoritmima 1, 2 i 3 (za prag sličnosti 30%).

C. Glavna petlja Algoritma 3

```
#=============================================================================
#         GLAVNA PETLJA ALGORITMA
#=============================================================================


#=============================================================================
# UČITAVANJE LISTE PROTEINA ZA REDUCIRANU MATRICU,OSTAVJANJE JEDINSTVENIH
# prva lista su proteinska obilježja, druga lista su indeksi susjeda + vlastiti indeks
# Prva lista [redni broj, pdb, duljina, TM, frekvencija]
#=============================================================================

sortiranje = [3,4] #izbor stupaca za sortiranje u petlji za protein
rezultati = open('Rezultati_analize_seta_final.txt','w')
for s in range(len(sortiranje)):
    for raz in range(0,10):
        ostavljeni = []
        zalihni = []
        red_mat = mat
        for iter in range (0,2500):
            if red_mat.shape == (0,4):
                print 'Završavam'
                break;
            for i in range(len(red_mat[:,0])):
                if red_mat[i,0] in zalihni or red_mat[i,0] in ostavljeni:
                    print 'Obrisani ili zalihni protein je ostao u matrici'
                    break;
            lista_proteina = []
            for i in range(len(red_mat[:,0])):
                indeksi = []
                for j in range(4,len(red_mat[i,:])):
                    if red_mat[i,j] == 1:
                        indeksi.append(red_mat[j-4,0])
                indeksi.remove(red_mat[i,0])
                if len(indeksi) == 0:
```

```
                    print 'indeksi u matrici su jednaki nula'
                    print red_mat
                    break;
                for sp in select_prot:
                    if red_mat[i,0] == sp[0]:
                        pk = sp[0:4]+[float("%.2f" % (float(sp[3])/len(indeksi)))]+[len(indeksi)]
                lista_proteina.append(pk+[indeksi])

#=============================================================================
#             # SORTIRANJE LISTE PROTEINA PO ŽELJENIM KRITERIJIMA RADI IZBORA RAZMATRANOG PRVOG
#             #0 - indeks, 1- pdb, 2-len, 3-tm, 4-tm/len, 5-susjedi
#=============================================================================

            lista_proteina = sorted(lista_proteina, key = itemgetter(4), reverse = True)
            #lista_proteina = sorted(lista_proteina, key = itemgetter(5))#, reverse = True)
            print 'Lista proteina ima elemenata: ', len(lista_proteina), 'a reducirana matrica:', red_mat.shape,
             '\n\n'
            if iter == 0:
                ulazna_lista = lista_proteina
            ###########################################################################
            prot_ostavljeni = [] #lista aostavljenih proteina u iteraciji uz razmatrani protein
            prot_zalihni = [] #Lista zalihnih proteina u iteraciji uz razmatrani protein
            protein = lista_proteina[0][0]#UZIMA SE U OBZIR PRVI IZ LISTE SA SORTIRANJEM
            #print 'RAZMATRA SE PROTEIN: ', protein, lista_proteina[0]

              #kontrolni parametri
            if protein in ostavljeni:
                print 'PAŽNJA, RAZMATRANI PROTEIN JE U OSTVALJENIM', protein
                continue;
            if protein in zalihni:
                print 'PAŽNJA, RAZMATRANI PROTEIN JE U ZALIHNIM', protein
                continue;

               #PROTEIN NEMA SUSJEDA, NITI SA SOBOM OPREZ
            if len(lista_proteina[0][6]) == 0:
                print 'Problem jer protein nema susjeda: '
                break;

        #=============================================================================
        #  ako razmatrani protein ima samo jednog susjeda
```

```
#=============================================================================
    if len(lista_proteina[0][6]) == 1:
        print lista_proteina[0]
        prot_ostavljeni.append(protein)
        prot_zalihni.append(lista_proteina[0][6][0])
        if lista_proteina[0][6][0] in prot_ostavljeni:
            print 'Ostavljeni 1'
        print 'Ostavljen je: ', protein, 'a izbačen je: ', lista_proteina[0][6]

#=============================================================================
#       TRAŽENJE ONIH SA ZAJEDNIČKIM SUSJEDIMA
#=============================================================================
    else:
        prot_ostavljeni.append(protein)
        for z in lista_proteina[0][6]:
            prot_zalihni.append(z)
            print 'Ostavljeni 2'
            if z in ostavljeni:
                print 'Protein iz liste je u ostavljenim, oprez'
                break;
        lista_susjeda = []
        for i,lp in enumerate(lista_proteina):


                #KONTROLNI PARAMETRI
            if lp[0] in ostavljeni:
                print i,'Protein je u ostavljenim', lp[0]
                continue;
            if lp[0] in zalihni:
                print i,'Protein je u zalihnim', lp[0]
                continue

                #petlja u kojoj se definira ntorka karakterističnih podataka (9-orka)
            if i > 0:
                presjek = set(lista_proteina[0][6]).intersection(lp[6])
                if protein not in lp[6] and len(presjek) > 0:
                      #ako lanci nisu slični i imaju zajedničkih susjeda
                    print 'Razmatra se protein sa zajedničkim susjedima: ', protein, lp[0]
                    if presjek == lp[6]:
                        prot_ostavljeni.append(lp[0])
```

```
                print '\n\n\nPresjek\n\n\n'
                print lista_proteina[0][6], lp[6]
                print '\n\n\nPresjek\n\n\n'
                continue;
            else:
                susout = (len(lp[6])-len(presjek))
                if susout == 0:
                    prot_ostavljeni.append(lp[0])
                else:
                    tmsus = float("%.2f" % (float(lp[3])/susout))
                    #print i, l[0], len(presjek), len(l[6]), l[6]
                    ntorka = lp
                    ntorka.insert(6,susout)
                    ntorka.insert(6,len(presjek))
                    ntorka.insert(5,tmsus)
                    lista_susjeda.append(ntorka)

#sortiranje se vrši po podatku iz entorke
sort = int(sortiranje[s])
granica = float(0.2+(0.2*raz))
     lista_susjeda = sorted(lista_susjeda, key=itemgetter(sort), reverse=True)

    #KRITERIJ SORTIRANJA
# printanje liste susjeda
for i,ls in enumerate(lista_susjeda):
    print i,'-ti susjed: ', ls

print 'Ostavljeni su:', prot_ostavljeni
print 'Zalihni su:', prot_zalihni

lista_podsusjeda = lista_susjeda#definira se lista za iteracije
for i,ls in enumerate(lista_susjeda):
    if lista_podsusjeda[0][sort] >= granica:
        continue;
    if ls[0] in zalihni:
            continue;
    if ls[0] in ostavljeni:
        print 'Oprez lanac je u ostavljenom', ls[0]
        continue;
         #POSTAVLJANJE GRANICE OSTAVLJANJA PROTEINSKIH LANACA, OMJER
```

```
            if lista_podsusjeda[0][sort] <= granica:
                if lista_podsusjeda[0][0] in prot_zalihni:
                    continue;
                prot_ostavljeni.append(lista_podsusjeda[0][0])
                for zs in lista_podsusjeda[0][9]: # zs - zalihni susjed
                    prot_zalihni.append(zs)
                    print 'Ostavljeni 3'
                    print 'Zalihni je: ', zs
                print 'Ostavljeni protein je obrisan: ', lista_podsusjeda[0][0]
                del lista_podsusjeda[0]
                for j,lps in enumerate(lista_podsusjeda):
                    if lps[0] in prot_ostavljeni or lps[0] in prot_zalihni:
                        print 'Obrisani iz podsusjeda su: ', lps[0]
                        del lps

                    else:
                        podpresjek = set(prot_zalihni).intersection(lps[9])
                        if podpresjek == lps[9]:
                            prot_ostavljeni.append(lps[j])
                            print '\n\n\nPresjek\n\n\n'
                            print lps[0], lps[9]
                            print '\n\n\nPresjek\n\n\n'
                            continue;
                        else:
                            podsusout = (len(lps[9])-len(podpresjek))
                            if podsusout == 0:
                                prot_ostavljeni.append(lps[0])
                            else:
                                lps[8] = podsusout
                                tmsusout = float("%.2f" % (float(lps[3])/podsusout))
                                lps[5] = tmsusout
            lista_podsusjeda = sorted(lista_podsusjeda, key=itemgetter(sort),       reverse=True) #KRITERIJ
     SORTIRANJA
    ostavljeni = ostavljeni + prot_ostavljeni
    ostavljeni = list(set(ostavljeni))
    zalihni = zalihni + prot_zalihni
    zalihni = list(set(zalihni))
#=============================================================================
#                lista_susjeda.append[l[0]]
#         print 'Susjedi proteinu su:', lista_susjeda
```

```
#================================================================================
    prot_brisati_r = []
    for po in prot_ostavljeni:
        for i,lp in enumerate(lista_proteina):
            if po == lp[0]:
                #print lp
                continue;
            del lp
    brisati_r = []
    for i in range(len(red_mat[:,0])):
        if red_mat[i,0] in ostavljeni:
            brisati_r.append(i)
            print 'Briše se protein',red_mat[i,0],'u retku', i
    for j in range(len(red_mat[:,0])):
        if red_mat[j,0] in zalihni:
            brisati_r.append(j)
            print 'Briše se protein',red_mat[j,0],'u retku', j
    brisati_c = [br+4 for br in brisati_r]
    red_mat = np.delete(red_mat,brisati_r,axis=0)
    red_mat = np.delete(red_mat,brisati_c,axis=1)

      #================================================================================
    #     izbacivanje ostavljenih i zalihnih lanaca, te jedinstvenih iz matrice
    #================================================================================

    if red_mat.shape == (0,4):
        print 'Završavam'
        continue;
    izbaciti_j = []
    freq_red = sum(red_mat[:,4:]) #suma po svim stupcima od  4. pa nadalje po cijelom stupcu
    red_mat[:,3] = freq_red #ubacivanje novih frekvencija
    for i,fr in enumerate(freq_red):
        if fr == 1:
            ostavljeni.append(red_mat[i,0])
            izbaciti_j.append(i)
    izbaciti_c = [ij+4 for ij in izbaciti_j]
    red_mat = np.delete(red_mat,izbaciti_j,axis=0)
    red_mat = np.delete(red_mat,izbaciti_c,axis=1)
ostavljeni_total = ostavljeni_ul + ostavljeni
zalihni_total = zalihni_ul + zalihni
```

D. Reprezentativni skup N234 s 234 lanca dobivenih primjenom Algoritma 3 na početni skup N1212 uz prag sličnosti 30%

PDB_KOD|DULJINA_SLIJEDA|BROJ_TM|REZOLUCIJA|EKSP.METODA|$Q_{2,rnd}$|BINOMNI_KOEF.|SEGMENTNI_KOEF.
PDB_SLIJED(*-SIMBOL DODAN KAKO BI SE REDOSLIJED AMINOKISELINA U MEMBRANI PREMA OPM-u PODUDARIO S REDOSLIJEDOM TIH AMINOKISELINA U SLIJEDU IZ PDB-a)
POLOŽAJI_TM_SEGMENATA

4yl3_A|148|4|1.41|X-RAY|0.54|93.27|15.3
****SLVMSSPALPAFLLCSTLLVIKMYVVAIITGQVRLRKKAFANPEDALRHGGPQYCRSDPDVERCLRAHRNDMETIYPFLFLGFVYSFLGPNPFVAWMHFLVFLVGRVAHTVAYLGKLR
APIRSVTYTLAQLPCASMALQILWEAARHL
1|13|36|2|68|90|3|97|119|4|124|149

5aez_A|486|11|1.47|X-RAY|0.5|331.74|57.64
MSGQFTGTGTGGDVFKVDLNEQFDRADMVWIGTASVLVWIMIPGVGLLYSGISRKKHALSLMWAALMAACVAAFQWFWWGYSLVFAHNGSVFLGTLQNFCLKDVLGAPSIVKTVPDILFCLY
QGMFAAVTAILMAGAGCERARLGPMMVFLFIWLTVVYCPIAYWTWGGNGWLVSLGALDFAGGGPVHENSGFAALAYSLWLGKRHDPVAKGKVPKYKPHSVSSIVMGTIFLWFGWYGFNGGST
GNSSMRSWYACVNTNLAAATGGLTWMLVDWFRTGGKWSTVGLCMGAIAGLVGITPAAGYVPVYTSVIFGIVPAIICNFAVDLKDLLQIDDGMDVWALHGVGGFVGNFMTGLFAADYVAMIDG
TEIDGGWMNHHWKQLGYQLAGSCAVAAWSFTVTSIILLAMDRIPFLRIRLHEDEEMLGTDLAQIGEYAYYADDDPETNPYVLEPIRSTTISQPLPHIDGVADGSSNNDSGEAKNHHHHHH
1|29|52|2|60|86|3|116|138|4|144|168|5|184|201|6|223|244|7|250|276|8|283|305|9|306|326|10|336|357|11|378|409

4zw9_A|518|12|1.5|X-RAY|0.51|352.88|62.51
MHHHHHHHHHHSGDEVDAGSGHMGTQKVTPALIFAITVATIGSFQFGYNTGVINAPEKIIKEFITKTLTDKGNAPPSEVLLTSLWSLSVAIFSVGGMIGSFSVGLFVNRFGRRNSMLIVNLL
AVTGGCFMGLCKVAKSVEMLILGRLVIGLFCGLCTGFVPMYIGEISPTALRGAFGTLNQLGIVVGILVAQIFGLEFILGSEELWPLLLGFTILPAILQSAALPFCPESPRFLLINRKEEENA
KQILQRLWGTQDVSQDIQEMKDESARMSQEKQVTVLELFRVSSYRQPIIISIVLQLSQQLSGINAVFYYSTGIFKDAGVQEPIYATIGAGVVNTIFTVVSLFLVERAGRRTLHMIGLGGMAF
CSTLMTVSLLLKDNYNGMSFVCIGAILVFVAFFEIGPGPIPWFIVAELFSQGPRPAAMAVAGCSNWTSNFLVGLLFPSAAHYLGAYVFIIFTGFLITFLAFTFFKVPETRGRTFEDITRAFE
GQAHGADRSGKDGVMEMNSIEPAKETTTNV
1|11|34|2|63|85|3|92|111|4|116|140|5|152|174|6|184|204|7|270|296|8|304|326|9|331|355|10|361|389|11|399|424|12|429|448

2xfn_A|520|1|1.6|X-RAY|0.96|51.18|6.23
MSNKCDVVVVGGGISGMAAAKLLHDSGLNVVVLEARDRVGGRTYTLRNQKVKYVDLGGSYVGPTQNRILRLAKELGLETYKVNEVERLIHHVKGKSYPFRGPFPPVWNPITYLDHNNFWRTM
DDMGREIPSDAPWKAPLAEEWDNMTMKELLDKLCWTESAKQLATLFVNLCVTAETHEVSALWFLWYVKQCGGTTRIISTTNGGQERKFVGGSGQVSERIMDLLGDRVKLERPVIYIDQTREN
VLVETLNHEMYEAKYVISAIPPTLGMKIHFNPPLPMMRNQMITRVPLGSVIKCIVYYKEPFWRKKDYCGTMIIDGEEAPVAYTLDDTKPEGNYAAIMGFILAHKARKLARLTKEERLKKLCE
LYAKVLGSLEALEPVHYEEKNWCEEQYSGGCYTTYFPPGILTQYGRVLRQPVDRIYFAGTETATHWSGYMEGAVEAGERAAREILHAMGKIPEDEIWQSEPESVDVPAQPITTTFLERHLPS
VPGLLRLIGLTTIFSATALGFLAHKRGLLVRV
1|489|499

1kqf_B|294|1|1.6|X-RAY|0.86|75.76|5.61
MAMETQDIIKRSATNSITPPSQVRDYKAEVAKLIDVSTCIGCKACQVACSEWNDIRDEVGHCVGVYDNPADLSAKSWTVMRFSETEQNGKLEWLIRKDGCMHCEDPGCLKACPSAGAIIQYA
NGIVDFQSENCIGCGYCIAGCPFNIPRLNKEDNRVYKCTLCVDRVSVGQEPACVKTCPTGAIHFGTKKEMLELAEQRVAKLKARGYEHAGVYNPEGVGGTHVMYVLHHADQPELYHGLPKDP
KIDTSVSLWKGALKPLAAAGFIATFAGLIFHYIGIGPNKEVDDDEEDHHE
1|256|277

4rp9_A|465|11|1.65|X-RAY|0.5|318.99|59.64
MEILYNIFTVFFNQVMTNAPLLLGIVTCLGYILLRKSVSVIIKGTIKTIIGFMLLQAGSGILTSTFKPVVAKMSEVYGINGAISDTYASMMATIDRMGDAYSWVGYAVLLALALNICYVLLR
RITGIRTIMLTGHIMFQQAGLIAVTLFIFGYSMWTTIICTAILVSLYWGITSNMMYKPTQEVTDGCGFSIGHQQQFASWIAYKVAPFLGKKEESVEDLKLPGWLNIFHDNIVSTAIVMTIFF
GAILLSFGIDTVQAMAGKVHWTVYILQTGFSFAVAIFIITQGVRMFVAELSEAFNGISQRLIPGAVLAIDCAAIYSFAPNAVVWGFMWGTIGQLIAVGILVACGSSILIIPGFIPMFFSNAT
IGVFANHFGGWRAALKICLVMGMIEIFGCVWAVKLTGMSAWMGMADWSILAPPMMQGFFSIGIAFMAVIIVIALAYMFFAGRALRAEEDAEKQLAEQSA
1|15|33|2|41|73|3|104|121|4|134|145|5|156|178|6|232|251|7|265|291|8|324|344|9|357|372|10|377|398|11|429|447

4o6y_A|230|6|1.7|X-RAY|0.51|153.66|25.87
MAVPVLGGFPIFMVVRVLGFIIAALVLTWTVHYRGGLALSSDNKDHIFNVHPVMMVIGLILFNGEAMLAYKSVQGTKNLKKLVHLTLQLTAFILSLIGVWAALKFHIDKGIENFYSLHSWLG
LACLFLFAFQWAAGFVTYWYPGGSRNSRASLMPWHVFLGISIYALALVTATTGILEKVTFLQVNQVITRYSTEAMLVNTMGVLILILGGFVILGVVTPVSGKDQVLTQ
1|12|31|2|48|69|3|81|103|4|117|138|5|156|179|6|196|217

4v1f_B|86|2|1.7|X-RAY|0.51|56.58|6.66
MELDPNALITAGALIGGGLIMGGGAIGAGIGDGIAGNALISGIARQPEAQGRLFTPFFITVGLVEAAYFINLAFMALFVFATPGLQ
1|14|37|2|57|80

2xtv_A|180|6|1.7|X-RAY|0.51|119.45|25.48
**AGPVTWVMMIACVVVFIAMQILGDQEVMLW
LAWPFDPTLKFEFWRYFTHALMHFSLMHILFNLLWWWYLGGAVEKRLGSGKLIVITLISALLSGYVQQKFSGPWFGGLTGVVYALMGYVWLRGERDPQSGIYLQRGLIIFALIWIVAGWFDL
FGMSMANGAHIAGLAVGLAMAFVDSLNA
1|95|114|2|148|163|3|171|192|4|201|213|5|227|242|6|251|268

2bs2_C|256|5|1.78|X-RAY|0.5|174.45|23.68
MTNESILESYSGVTPERKKSRMPAKLDWWQSATGLFLGLFMIGHMFFVSTILLGDNVMLWVTKKFELDFIFEGGKPIVVSFLAAFVFAVFIAHAFLAMRKFPINYRQYLTFKTHKDLMRHGD
TTLWWIQAMTGFAMFFLGSVHLYIMMTQPQTIGPVSSSFRMVSEWMWPLYLVLLFAVELHGSVGLYRLAVKWGWFDGETPDKTRANLKKLKTLMSAFLIVLGLLTFGAYVKKGLEQTDPNID
YKYFDYKRTHEE
1|28|53|2|76|98|3|128|154|4|162|188|5|208|232

3s8g_A|569|13|1.8|X-RAY|0.5|388.34|66.82
MHHHHHHHAVRASEISRVYEAYPEKKATLYFLVLGFLALIVGSLFGPFQALNYGNVDAYPLLKRLLPFVQSYYQGLTLHGVLNAIVFTQLFAQAIMVYLPARELNMRPNMGLMWLSWWMAFI
GLVVFALPLLANEATVLYTFYPPLKGHWAFYLGASVFVLSTWVSIYIVLDLWRRWKAANPGKVTPLVTYMAVVFWLMWFLASLGLVLEAVLFLLPWSFGLVEGVDPLVARTLFWWTGHPIVY
FWLLPAYAIIYTILPKQAGGKLVSDPMARLAFLLFLLLSTPVGFHHQFADPGIDPTWKMIHSVLTLFVAVPSLMTAFTVAASLEFAGRLRGGRGLFGWIRALPWDNPAFVAPVLGLLGFIPG
GAGGIVNASFTLDYVVHNTAWVPGHFHLQVASLVTLTAMGSLYWLLPNLTGKPISDAQRRLGLAVVWLWFLGMMIMAVGLHWAGLLNVPRRAYIAQVPDAYPHAAVPMVFNVLAGIVLLVAL
LLFIYGLFSVLLSRERKPELAEAPLPFAEVISGPEDRRLVLAMDRIGFWFAVAAILVVLAYGPTLVQLFGHLNPVPGWRLW
1|21|47|2|67|89|3|104|129|4|143|162|5|185|210|6|223|245|7|267|285|8|292|314|9|347|372|10|379|402|11|420|444|12|465|490|13|
527|550

2dyr_G|85|1|1.8|X-RAY|0.62|46.29|4.16
ASAAKGDHGGTGARTWRFLTFGLALPSVALCTLNSWLHSGHRERPAFIPYHHLRIRTKPFSWGDGNHTFFHNPRVNPLPTGYEKP
1|16|37

2dyr_L|47|1|1.8|X-RAY|0.5|30.33|3.26
SHYEEGPGKNIPFSVENKWRLLAMMTLFFGSGFAAPFFIVRHQLLKK
1|22|43

2dyr_J|59|1|1.8|X-RAY|0.54|36.19|3.66
FENRVAEKQKLFQEDNGLPVHLKGGATDNILYRVTMTLCLGGTLYSLYCLGWASFPHKK
1|34|54

2dyr_I|73|1|1.8|X-RAY|0.59|41.53|3.97
STALAKPQMRGLLARRLRFHIVGAFMVSLGFATFYKFAVAEKRKKAYADFYRNYDSMKDFEEMRKAGIFQSAK
1|16|36

3s8g_B|168|1|1.8|X-RAY|0.76|64.67|4.98
MVDEHKAHKAILAYEKGWLAFSLAMLFVFIALIAYTLATHTAGVIPAGKLERVDPTTVRQEGPWADPAQAVVQTGPNQYTVYVLAFAFGYQPNPIEVPQGAEIVFKITSPDVIHGFHVEGTN
INVEVLPGEVSTVRYTFKRPGEYRIICNQYCGLGHQNMFGTIVVKE
1|15|37

2dyr_C|261|7|1.8|X-RAY|0.5|177.47|31.42
MTHQTHAYHMVNPSPWPLTGALSALLMTSGLTMWFHFNSMTLLMIGLTTNMLTMYQWWRDVIRESTFQGHHTPAVQKGLRYGMILFIISEVLFFTGFFWAFYHSSLAPTPELGGCWPPTGIH
PLNPLEVPLLNTSVLLASGVSITWAHHSLMEGDRKHMLQALFITITLGVYFTLLQASEYYEAPFTISDGVYGSTFFVATGFHGLHVIIGSTFLIVCFFRQLKFHFTSNHHFGFEAAAWYWHF
VDVVWLFLYVSIYWWGS
1|17|35|2|41|55|3|82|100|4|126|146|5|161|181|6|197|217|7|237|258

2j8c_H|260|1|1.87|X-RAY|0.86|68.13|5.48
MVGVTAFGNFDLASLAIYSFWIFLAGLIYYLQTENMREGYPLENEDGTPAANQGPFPLPKPKTFILPHGRGTLTVPGPESEDRPIALARTAVSEGFPHAPTGDPMKDGVGPASWVARRDLPE
LDGHGHNKIKPMKAAAGFHVSAGKNPIGLPVRGCDLEIAGKVVDIWVDIPEQMARFLEVELKDGSTRLLPMQMVKVQSNRVHVNALSSDLFAGIPTIKSPTEVTLLEEDKICGYVAGGLMYA
APKRKSVVAAMLAEYA
1|13|32

2j8c_L|281|5|1.87|X-RAY|0.52|186.72|22.5
ALLSFERKYRVPGGTLVGGNLFDFWVGPFYVGFFGVATFFFAALGIILIAWSAVLQGTWNPQLISVYPPALEYGLGGAPLAKGGLWQIITICATGAFVSWALREVEICRKLGIGYHIPFAFA
FAILAYLTLVLFRPVMMGAWGYAFPYGIWTHLDWVSNTGYTYGNFHYNPAHMIAISFFFTNALALALHGALVLSAANPEKGKEMRTPDHEDTFFRDLVGYSIGTLGIHRLGLLLSLSAVFFS
ALCMIITGTIWFDQWVDWWQWWVKLPWWANIPGGING
1|33|54|2|84|106|3|116|138|4|170|192|5|229|251

3cx5_I|65|1|1.9|X-RAY|0.63|34.1|3.91
*SFSSLYKTFFKRNAVFVGTIFAGAFVFQTVFDTAITSWYENHNKGKLWKDVKARIAAGDGDDDDE
1|18|33

3wu2_t|32|1|1.9|X-RAY|0.53|19.24|2.56
METITYVFIFACIIALFFFAIFFREPPRITKK
1|4|23

4xu4_A|210|6|1.9|X-RAY|0.51|140.22|24.68
MRLRISEAVVLFLLGAVAALIGDHSHVVTGTTVYHTDAVPFVWSSPFWFPILVGAATASLAELRLHLPAPRDGVTARQALGGVAAVVGTYVTTALVHAFPVVPVTALVCAAAAITWCVLGDG
PGAACGVVIAVIGPAVEIALVQLGVFAYHPDSDGLFGVAPFLAPLYFAFGVVAALLGELAVARRPQLGPPVCDTVSRGPGAGHHHHHH
1|5|27|2|47|64|3|76|97|4|101|118|5|124|145|6|162|179

2qts_A|438|2|1.9|X-RAY|0.8|152.72|11.93
**********STLHGISHIFSYERLSLKRVVWALCFMGSLALLALVCTNRIQYYFLYPHVTKLDEVAATRLTFPAVTFCNLNEFRFSRVTKNDLYHAGELLALLNNR
YEIPDTQTADEKQLEILQDKANFRNFKPKPFNMLEFYDRAGHDIREMLLSCFFRGEQCSPEDFKVVFTRYGKCYTFNAGQDGKPRLITMKGGTGNGLEIMLDIQQDEYLPVWGETDETSFEA
GIKVQIHSQDEPPLIDQLGFGVAPGFQTFVSCQEQRLIYLPPPWGDCKATTGDSEFYDTYSITACRIDCETRYLVENCNCRMVHMPGDAPYCTPEQYKECADPALDFLVEKDNEYCVCEMPC
NVTRYGKELSMVKIPSKASAKYLAKKYNKSEQYIGENILVLDIFFEALNYETIEQKKAYEVAGLLGDIGGQMGLFIGASILTVLELFDYAYEVIKHR
1|45|67|2|427|453

5jeq_A|236|2|1.9|X-RAY|0.68|115.08|10.5
MIVKRPVSASLARAFFYIVLLSILSTGIALLTLASSLRDAEAINIAGSLKMQSYRLGYDLQSGSPQLNAHRQLFQQALHSPVLTNLNVWYVPEAVKTRYAHLNANWLEMNNRLSKGDLPWYQ
ANINNYVNQIDLFVLALQHYAERKMLLVVAISLAGGIGIFTLVFFTLRRIRHQVVAPLNQLVTASQRIEHGQFDSPPLDTNLPNELGLLAKTFNQMSSELHKLYRSLEHHHHHH
1|12|34|2|145|168

4dx5_B|1057|12|1.9|X-RAY|0.64|573.5|74.77
MPNFFIDRPIFAWVIAIIIMLAGGLAILKLPVAQYPTIAPPAVTISASYPGADAKTVQDTVTQVIEQNMNGIDNLMYMSSNSDSTGTVQITLTFESGTDADIAQVQVQNKLQLAMPLLPQEV
QQQGVSVEKSSSSFLMVVGVINTDGTMTQEDISDYVAANMKDAISRTSGVGDVQLFGSQYAMRIWMNPNELNKFQLTPVDVITAIKAQNAQVAAGQLGGTPPVKGQQLNASIIAQTRLTSTE
EFGKILLKVNQDGSRVLLRDVAKIELGGENYDIIAEFNGQPASGLGIKLATGANALDTAAAIRAELAKMEPFFPSGLKIVYPYDTTPFVKISIHEVVKTLVEAIILVFLVMYLFLQNFRATL
IPTIAVPVVLLGTFAVLAAFGFSINTLTMFGMVLAIGLLVDDAIVVVENVERVMAEEGLPPKEATRKSMGQIQGALVGIAMVLSAVFVPMAFFGGSTGAIYRQFSITIVSAMALSVLVALIL
TPALCATMLKPIAKGDHGEGKKGFFGWFNRMFEKSTHHYTDSVGGILRSTGRYLVLYLIIVVGMAYLFVRLPSSFLPDEDQGVFMTMVQLPAGATQERTQKVLNEVTHYYLTKEKNNVESVF
AVNGFGFAGRGQNTGIAFVSLKDWADRPGEENKVEAITMRATRAFSQIKDAMVFAFNLPAIVELGTATGFDFELIDQAGLGHEKLTQARNQLLAEAAKHPDMLTSVRPNGLEDTPQFKIDID
QEKAQALGVSINDINTTLGAAWGGSYVNDFIDRGRVKKVYVMSEAKYRMLPDDIGDWYVRAADGQMVPFSAFSSSRWEYGSPRLERYNGLPSMEILGQAAPGKSTGEAMELMEQLASKLPTG
VGYDWTGMSYQERLSGNQAPSLYAISLIVVFLCLAALYESWSIPFSVMLVVPLGVIGALLAATFRGLTNDVYFQVGLLTTIGLSAKNAILIVEFAKDLMDKEGKGLIEATLDAVRMRLRPIL
MTSLAFILGVMPLVISTGAGSGAQNAVGTGVMGGMVTATVLAIFFVPVFFVVVRRRFSRKNEDIEHSHTVDHHLEHHHHHH
1|8|27|2|339|359|3|365|383|4|397|414|5|438|456|6|471|493|7|538|556|8|874|894|9|896|918|10|927|947|11|971|991|12|1003|1026

3wu2_b|504|6|1.9|X-RAY|0.6|294.56|33.36
*GLPWYRVHTVLINDPGRLIAAHLMHTALVAGWAGSMALYELATFDPSDPVLNPMWRQGMFVLPFMARLGVTGSWSGWSITGETGIDPGFWSFEGVALAHIVLSGLLFLAACWHWVYWDLEL
FRDPRTGEPALDLPKMFGIHLFLAGLLCFGFGAFHLTGLFGPGMWVSDPYGLTGSVQPVAPEWGPDGFNPYNPGGVVAHHIAAGIVGIIAGLFHILVRPPQRLYKALRMGNIETVLSSSIAA
VFFAAFVVAGTMWYGSATTPIELFGPTRYQWDSSYFQQEINRRVQASLASGATLEEAWSAIPEKLAFYDYIGNNPAKGGLFRTGPMNKGDGIAQAWKGHAVFRNKEGEELFVRRMPAFFESF
PVILTDKNGVVKADIPFRRAESKYSFEQQGVTVSFYGGELNGQTFTDPPTVKSYARKAIFGEIFEFDTETLNSDGIFRTSPRGWFTFAHAVFALLFFFGHIWHGARTLFRDVFSGIDPELSP
EQVEWGFYQKVGDVTTR
1|17|39|2|95|116|3|135|159|4|197|218|5|234|256|6|450|474

1ors_C|132|5|1.9|X-RAY|0.62|72.77|18.05
**********DVMEHPLVELGVSYAALLSVIVVVVEYTMQLSGEYLVRLYLVDLILVIILWADYAYRAYKSGDPAGYVKKTLYEIPALVPAGLLALIEGHLAGLGLFRLVRLL
RFLRILLIISRGSKFLSAIADAADKLVPR
1|25|46|2|55|78|3|86|97|4|100|107|5|117|148

3cx5_H|93|1|1.9|X-RAY|0.67|44.81|4.32
*GPPSGKTYMGWWGHMGGPKQKGITSYAVSPYAQKPLQGIFHNAVFNSFRRFKSQFLYVLIPAGIYWYWWKNGNEYNEFLYSKAGREELERVNV
1|55|73

3cx5_D|248|1|1.9|X-RAY|0.85|67.15|5.43
**MTAAEHGLHAPAYAWSHNGPFETFDHASIRRGYQVYREVCAACHSLDRVAWRTLVGVSHTN
EEVRNMAEEFEYDDEPDEQGNPKKRPGKLSDYIPGPYPNEQAARAANQGALPPDLSLIVKARHGGCDYIFSLLTGYPDEPPAGVALPPGSNYNPYFPGGSIAMARVLFDDMVEYEDGTPATT
SQMAKDVTTFLNWCAEPEHDERKRLGLKTVIILSSLYLLSIWVKKFKWAGIKTRKFVFNPPKPRK
1|270|289

1h2s_B|60|2|1.93|X-RAY|0.56|35.97|6.14
**********************GAVFIFVGALTVLFGAIAYGEVTAAAATGDAAAVQEAAVSAILGLIILLGINLGLVAATL
1|24|41|2|60|81

2hil_I|158|1|12.5|E-MIC|0.78|57.66|4.93
FTLIELMIVIAIVGILAAVALPAYQDYTARAQVSEAILLAEGQKSAVTEYYLNHGKWPENNTSAGVASSPTDIKGKYVKEVEVKNGVVTATMLSSGVNNEIKGKKLSLWARRENGSVKWFCG
QPVTRTDDDTVADAKDGKEIDTKHLPSTCRDNFDAK
1|5|24

1xio_A|261|7|2.0|X-RAY|0.52|173.66|31.58
MNLESLLHWIYVAGMTIGALHFWSLSRNPRGVPQYEYLVAMFIPIWSGLAYMAMAIDQGKVEAAGQIAHYARYIDWMVTTPLLLLSLSWTAMQFIKKDWTLIGFLMSTQIVVITSGLIADLS
ERDWVRYLWYICGVCAFLIILWGIWNPLRAKTRTQSSELANLYDKLVTYFTVLWIGYPIVWIIGPSGFGWINQTIDTFLFCLLPFFSKVGFSFLDLHGLRNLNDSRQTTGDRFAENTLQFVE
NITLFANSRRQQSRRRV
1|3|26|2|34|56|3|70|89|4|99|121|5|126|147|6|168|185|7|195|218

4x5m_A|100|3|2.0|X-RAY|0.55|61.63|10.75
MDTILLTGLFAAFFTTFAFAPQSIKTIRTRNTEGISVVMYIMFLTGVISWIAYGIMRSDFAVLIANIVTLFLAAPVLVITLINRRKKHVLESSGENLYFQ
1|4|24|2|37|58|3|60|82

1nkz_B|41|1|2.0|X-RAY|0.5|26.22|3.14
ATLTAEQSEELHKYVIDGTRVFLGLALVAHFLAFSATPWLH
1|18|36

1z98_A|281|7|2.1|X-RAY|0.5|190.62|33.6
MSKEVSEEAQAHQHGKDYVDPPPAPFFDLGELKLWSFWRAAIAEFIATLLFLYITVATVIGHSKETVVCGSVGLLGIAWAFGGMIFVLVYCTAGISGGHINPAVTFGLFLARKVSLLRALVY
MIAQCLGAICGVGLVKAFMKGPYNQFGGGANSVALGYNKGTALGAEIIGTFVLVYTVFSATDPKRSARDSHVPILAPLPIGFAVFMVHLATIPITGTGINPARSFGAAVIFNSNKVWDDQWI
FWVGPFIGAAVAAAYHQYVLRAAAIKALGSFRSNPTN
1|37|58|2|75|92|3|102|111|4|116|137|5|164|182|6|200|214|7|242|263

1lgh_A|56|1|2.4|X-RAY|0.54|34.3|3.61
SNPKDDYKIWLVINPSTWLPVIWIVATVVAIAVHAAVLAAPGFNWIALGAAKSAAK
1|20|39

5hwy_A|300|10|2.1|X-RAY|0.59|178.55|44.68
MVILGVGYFLLGLILLYYGSDWFVLGSERIARHFNVSNFVIGATVMAIGTSLPEILTSAYASYMHAPGISIGNAIGSCICNIGLVLGLSAIISPIIVDKNLQKNILVYLLFVIFAAVIGIDG
FSWIDGVVLLILFIIYLRWTVKNGSAEIEENNDKNNPSVVFSLVLLIIGLIGVLVGAELFVDGAKKIALALDISDKVIGFTLVAFGTSLPELMVSLAAAKRNLGGMVLGNVIGSNIADIGGA
LAVGSLFMHLPAENVQMAVLVIMSLLLYLFAKYSKIGRWQGILFLALYIIAIASLR
1|3|24|2|39|61|3|68|92|4|103|121|5|125|143|6|165|190|7|200|220|8|228|251|9|258|274|10|283|298

4u9n_A|178|5|2.2|X-RAY|0.57|108.1|19.69
****
****
**********LVYSEAGPVALWLARVRWLVILILTGMVTSSILQGFESVLEAVTALAFYVPVLLGTGGNTGNQSATLIIRALATRDLDLRDWRRVFLKEMGVGLLL
GLTLSFLLVGKVYWDGHPLLLPVVGVSLVLIVFFANLVGAMLPFLLRRLGVDPALVSNPLVATLSDVTGLLIYLSVARLLLE
1|284|306|2|316|337|3|352|381|4|386|409|5|421|443

3n5k_A|994|10|2.2|X-RAY|0.66|512.98|65.29
MEAAHSKSTEECLAYFGVSETTGLTPDQVKRHLEKYGHNELPAEEGKSLWELVIEQFEDLLVRILLLAACISFVLAWFEEGEETITAFVEPFVILLILIANAIVGVWQERNAENAIEALKEY
EPEMGKVYRADRKSVQRIKARDIVPGDIVEVAVGDKVPADIRILSIKSTTLRVDQSILTGESVSVIKHTEPVPDPRAVNQDKKNMLFSGTNIAAGKALGIVATTGVSTEIGKIRDQMAATEQ
DKTPLQQKLDEFGEQLSKVISLICVAVWLINIGHFNDPVHGGSWIRGAIYYFKIAVALAVAAIPEGLPAVITTCLALGTRRMAKKNAIVRSLPSVETLGCTSVICSDKTGTLTTNQMSVCKM
FIIDKVDGDFCSLNEFSITGSTYAPEGEVLKNDKPIRSGQFDGLVELATICALCNDSSLDFNETKGVYEKVGEATETALTTLVEKMNVFNTEVRNLSKVERANACNSVIRQLMKKEFTLEFS
RDRKSMSVYCSPAKSSRAAVGNKMFVKGAPEGVIDRCNYVRVGTTRVPMTGPVKEKILSVIKEWGTGRDTLRCLALATRDTPPKREEMVLDDSSRFMEYETDLTFVGVVGMLDPPRKEVMGS
IQLCRDAGIRVIMITGDNKGTAIAICRRIGIFGENEEVADRAYTGREFDDLPLAEQREACRRACCFARVEPSHKSKIVEYLQSYDEITAMTGDGVNDAPALKKAEIGIAMGSGTAVAKTASE
MVLADDNFSTIVAAVEEGRAIYNNMKQFIRYLISSNVGEVVCIFLTAALGLPEALIPVQLLWVNLVTDGLPATALGFNPPDLDIMDRPPRSPKEPLISGWLFFRYMAIGGYVGAATVGAAAW
WFMYAEDGPGVTYHQLTHFMQCTEDHPHFEGLDCEIFEAPEPMTMALSVLVTIEMCNALNSLSENQSLMRMPPWVNIWLLGSICLSMSLHFLILYVDPLPMIFKLKALDLTQWLMVLKISLP
VIGLDEILKFIARNYLEG
1|59|78|2|85|104|3|256|280|4|291|314|5|759|781|6|789|807|7|832|853|8|896|916|9|933|950|10|967|987

5j4i_A|453|12|2.21|X-RAY|0.54|293.84|59.25
MSSDADAHKVGLIPVTLMVSGNIMGSGVFLLPANLASTGGIAIYGWLVTIIGALGLSMVYAKMSFLDPSPGGSYAYARRCFGPFLGYQTNVLYWLACWIGNIAMVVIGVGYLSYFFPILKDP
LVLTITCVVVLWIFVLLNIVGPKMITRVQAVATVLALIPIVGIAVFGWFWFRGETYMAAWNVSGLGTFGAIQSTLNVTLWSFIGVESASVAAGVVKNPKRNVPIATIGGVLIAAVCYVLSTT
AIMGMIPNAALRVSASPFGDAARMALGDTAGAIVSFCAAAGCLGSLGGWTLLAGQTAKAAADDGLFPPIFARVNKAGTPVAGLIIVGILMTIFQLSSISPNATKEFGLVSSVSVIFTLVPYL
YTCAALLLLGHGHFGKARPAYLAVTTIAFLYCIWAVVGSGAKEVMWSFVTLMVITAMYALNYNRLHKNPYPLDAPISKDLELEVLFQ
1|12|32|2|40|61|3|84|113|4|124|145|5|146|173|6|189|213|7|224|247|8|273|301|9|324|343|10|347|370|11|385|408|12|409|427

2j7a_I|159|1|2.3|X-RAY|0.78|57.8|4.94
MSEEKSRNGPARLKLVLGGATLGVVALATVAFGMKYTDQRPFCTSCHIMNPVGVTHKLSGHANISCNDCHAPHNLLAKLPFKAIAGARDVYMNTLGHPGDLILAGMETKEVVNANCKACHTM
TNVEVASMEAKKYCTDCHRNVQHMRMKPISTREVADE
1|15|34

2zxe_B|305|1|2.4|X-RAY|0.83|93.26|5.62
MARGKSKETDGGWKKFLWDSEKKEFLGRTGSSWFKIFLFYLIFYGCLAGIFIGTIQVLLLTLSDFEPKYQDRVAPPGLSHAPYAIKTEISFSISNPKSYESFVKSMHKLMDLYNESSQAGNS
PFEDCSDTPADYIKRGDLDDSQGQKKACRFSRMWLKNCSGLDDTTYGYAEGKPCVVAKLNRIIGFYPKPLKNTTDLPEELQANYNQYVLPLRCAAKREEDREKIGSIEYFGLGGYAGFPLQY
YPYYGKRLQKKYLQPLLAIQFTNLTQNMELRIECKVYGENIDYSEKDRFRGRFEVKIEVKS
1|33|61

4ezc_A|384|10|2.36|X-RAY|0.5|261.95|51.06
MDDNPTAVKLDQGGNQAPQGRGRRCLPKALGYITGDMKEFANWLKDKPQALQFVDWVLRGISQVVFVSNPISGILILVGLLVQNPWCALNGCVGTVVSTLTALLLSQDRSAITAGLQGYNAT
LVGILMAIYSDKGNYFWWLLFPVSAMSMTCPVFSSALNSVLSKWDLPVFTLPFNMALSMYLSATGHYNPFFPSTLITPVTSVPNVTWPDLSALQLLKSLPVGVGQIYGCDNPWTGGIFLGAI
LLSSPLMCLHAAIGSLLGIIAGLSLSAPFEDIYAGLWGFNSSLACIAIGGTFMALTWQTHLLALACALFTAYLGASMSHVMAVVGLPSGTWPFCLATLLFLLLTTKNPNIYKMPISKVTYPE
ENRIFYLQSRKRTVQGPL
1|71|83|2|85|105|3|116|131|4|139|163|5|173|186|6|234|247|7|249|269|8|280|293|9|301|328|10|335|346

2yev_A|791|19|2.36|X-RAY|0.51|539.01|100.58
MAITAKPKAGVWAVLWDLLTTVDHKKIGLMYTATAFFAFALAGVFSLLIRTQLAVPNNQFLTGEQYNQILTLHGATMLFFFIIQAGLTGFGNFVVPLMLGARDVALPRVNAFSYWAFLGAIV
LALMSYFFPGGAPSVGWTFYYPFSAQSESGVDFYLAAILLLGFSSLLGNANFVATIYNLRAQGMSLWKMPIYVWSVFAASVLNLFSLAGLTAATLLVLLERKIGLSWFNPAVGGDPVLFQQF
FWFYSHPTVYVMLLPYLGILAEVASTFARKPLFGYRQMVWAQMGIVVLGTMVWAHHMFTVGESTLFQIAFAFFTALIAVPTGVKLFNIIGTLWGGKLQMKTPLYWVLGFIFNFLLGGITGVM
LSMTPLDYQFHDSYFVVAHFHNVLMAGSGFGAFAGLYYWWPKMTGRMYDERLGRLHFWLFLVGYLLTFLPQYALGYLGMPRRYYTYNADIAGWPELNLLSTIGAYILGLGGLVWIYTMWKSL
RSGPKAPDNPWGGYTLEWLTASPPKAHNFDVKLPTEFPSERPLYDWKKKGVELKPEDPAHIHLPNSSFWPFYSAATLFAFFVAVAALPVPNVWMWVFLALFAYGLVRWALEDEYSHPVEHHT
VTGKSNAWMGMAWFIVSEVGLFAILIAGYLYLRLSGAATPPEERPALWLALLNTFLLVSSSFTVHFAHHDLRRGRFNPFRFGLLVTIILGVLFFLVQSWEFYQFYHHSSWQENLWTAAFFTI
VGLHGLHVVIGGFGLILAYLQALRGKITLHNHGTLEAASMYWHLVDAVWLVIVTIFYVW
1|27|53|2|66|92|3|109|130|4|153|175|5|197|221|6|240|264|7|281|302|8|309|333|9|348|370|10|380|403|11|418|442|12|461|485|13|
556|574|14|580|597|15|620|641|16|657|677|17|691|711|18|728|752|19|766|789

2nq2_B|337|10|2.4|X-RAY|0.51|227.45|46.33
MQPDSYPKILFGLTLLLVITAVISLGIGRYSLSVPQIGQILWAKATALEIDPVQQQVIFQVRLPRILTALCVGAGLALSGVVLQGIFRNPLVNPHIIGVTSGSAFGGTLAIFFGFSLYGLFT
STILFGFGTLALVFLFSFKFNQRSLLMLILIGMILSGLFSALVSLLQYISDTEEKLPSIVFWLMGSFATSNWEKLLFFFVPFLLCSSILLSLSWRLNLLSLDEKEAKALGVKMAPLRWLVIF
LSGSLVACQVAISGSIGWVGLIIPHLSRMLVGANHQSLLPCTMLVGATYMLLVDNVARSLSDAEIPISILTALIGAPLFGVLVYKLKRGGMNE
1|7|26|2|62|84|3|99|111|4|117|139|5|148|170|6|194|214|7|240|257|8|263|271|9|282|304|10|311|328

3wmg_A|621|6|2.4|X-RAY|0.62|352.08|36.14
**********ASGPESAYTTGVTARRIFALAWSSSATMIV
IGFIASILEGATLPAFAIVFGRMFQVFTKSKSQIEGETWKYSVGFVGIGVFEFIVAGSRTALFGIASERLARDLRVAAFSNLVEQDVTYFDRRKAGELGGKLNNDVQVIQYSFSKLGAVLFN
LAQCVVGIIVAFIFAPALTGVLIALSPLVVLAVVVQMIEMSGNTKRSSEAYASAGSVAAEVFSNIRTTKAFEAERYETQRYGSKLDPLYRLGRRRYISDGLFFGLSMLVIFCVYALALWWGG
QLIARGSLNLGNLLTAFFSAILGFMGVGQAAQVWPDVTRGLGAGGELFAMIDRVPQYRRPDPGAEVVTQPLVLKQGIVFENVHFRYPTRMNVEVLRGISLTIPNGKTVAIVGGSGAGKSTII
QLLMRFYDIEPQGGGLLLFDGTPAWNYDFHALRSQIGLVSQEPVLFSGTIRDNILYGKRDATDEEVIQALREANAYSFVMALPDGLDTEVGERGLALSGGQKQRIAIARAILKHPTLLCLDE
STSALDAESEALVQEALDRMMASDGVTSVVIAHRLSTVARADLILVMQDGVVVEQGNHSELMALGPSGFYYQLVEKQLASGDMSAASGRDYKDDDDKHHHHHH
1|119|149|2|156|185|3|237|261|4|262|284|5|340|366|6|376|400

5jlc_A|515|1|2.4|X-RAY|0.92|88.45|6.2
**********LSYFQALPLAQRVSIMVALPFVYTITWQLLYSLRKDRPPLVFYWIPWVGSAIPYGTKPYEFFEDCQKKYGDIFSFMLLGRIMTVYLGPKGHEFIFNAKLADVS
AEAAYSHLTTPVFGKGVIYDCPNHRLMEQKKFVKGALTKEAFVRYVPLIAEEIYKYFRNSKNFKINENNSGIVDVMVSQPEMTIFTASRSLLGKEMRDKLDTDFAYLYSDLDKGFTPINFVF
PNLPLEHYRKRDHAQQAISGTYMSLIKERREKNDIQNRDLIDELMKNSTYKDGTKMTDQEIANLLIGVLMGGQHTSAATSAWCLLHLAERPDVQEELYQEQMRVLNNDTKELTYDDLQNMPL
LNQMIKETLRLHHPLHSLFRKVMRDVAIPNTSYVVPRDYHVLVSPGYTHLQEEFFPKPNEFNIHRWDGDAASSSAAGGDEVDYGFGAISKGVSSPYLPFGGGRHRCIGELFAYCQLGVLMSI
FIRTMKWRYPTEGETVPPSDFTSMVTLPTAPAKIYWEKRHPEQKYG
1|28|49

5lwy_A|292|7|2.4|X-RAY|0.51|197.12|32.39
GGSEFEGRWRVIPHDVLPDWLKDNDFLLHGHRPPMPSFRACFKSIFRIHTETGNIWTHLLGCVFFLCLGIFYMFRPNISFVAPLQEKVVFGLFFLGAILCLSFSWLFHTVYCHSEGVSRLFS
KLDYSGIALLIMGSFVPWLYYSFYCNPQPCFIYLIVICVLGIAAIIVSQWDMFATPQYRGVRAGVFLGLGLSGIIPTLHYVISEGFLKAATIGQIGWLMLMASLYITGAALYAARIPERFFP
GKCDIWFHSHQLFHIFVVAGAFVHFHGVSNLQEFRFMIGGGCSEEDAL**********
****
1|145|168|2|179|204|3|214|236|4|244|265|5|275|298|6|309|332|7|347|367

1zoy_C|140|3|2.4|X-RAY|0.5|94.29|12.05
***LGTTAKEEMERFWNKNLGSNRPLSPHITIYRWSLPMAMSICHRGTGIALSAGVSLFGLSALLLPGNFESHLELVKSLCLGPTLIYTAKFGIVFPLMYHTWNGIRHLIWDLGKGLTIPQL
TQSGVVVLILTVLSSVGLAAM
1|40|65|2|88|113|3|123|142

5hlb_A|747|1|2.42|X-RAY|0.95|82.49|6.59
**********KPRGKRGWLWLLLKLAIVFAVLIAIYGVYLDQKIRSRIDGKVWQLPAAVYGRMVNLEPDMTISKN
EMVKLLEATQYRQVSKMTRPGEFTVQANSIEMIRRPFDFPDSKEGQVRARLTFDGDHLATIVNMENNRQFGFFRLDPRLITMISSPNGEQRLFVPRSGFPDLLVDTLLATEDRHFYEHDGIS
LYSIGRAVLANLTAGRTVQGASTLTQQLVKNLFLSSERSYWRKANEAYMALIMDARYSKDRILELYMNEVYLGQSGDNEIRGFPLASLYYFGRPVEELSLDQQALLVGMVKGASIYNPWRNP
KLALERRNLVLRLLQQQQIIDQELYDMLSARPLGVQPRGGVISPQPAFMQLVRQELQAKLGDKVKDLSGVKIFTTFDSVAQDAAEKAAVEGIPALKKQRKLSDLETAIVVVDRFSGEVRAMV
GGSEPQFAGYNRAMQARRSIGSLAKPATYLTALSQPKIYRLNTWIADAPIALRQPNGQVWSPQNDDRRYSESGRVMLVDALTRSMNVPTVNLGMALGLPAVTETWIKLGVPKDQLHPVPAML
LGALNLTPIEVAQAFQTIASGGNRAPLSALRSVIAEDGKVLYQSFPQAERAVPAQAAYLTLWTMQQVVQRGTGRQLGAKYPNLHLAGKTGTTNNNVDTWFAGIDGSTVTITWVGRDNNQPTK
LYGASGAMSIYQRYLANQTPTPLNLVPPEDIADMGVDYDGNFVCSGGMRILPVWTSDPQSLCQQSEMQQQPS
1|69|86

2z73_A|448|7|2.5|X-RAY|0.53|295.1|36.96
MGRDLRDNETWWYNPSIVVHPHWREFDQVPDAVYYSLGIFIGICGIIGCGGNGIVIYLFTKTKSLQTPANMFIINLAFSDFTFSLVNGFPLMTISCFLKKWIFGFAACKVYGFIGGIFGFMS
IMTMAMISIDRYNVIGRPMAASKKMSHRRAFIMIIFVWLWSVLWAIGPIFGWGAYTLEGVLCNCSFDYISRDSTTRSNILCMFILGFFGPILIIFFCYFNIVMSVSNHEKEMAAMAKRLNAK
ELRKAQAGANAEMRLAKISIVIVSQFLLSWSPYAVVALLAQFGPLEWVTPYAAQLPVMFAKASAIHNPMIYSVSHPKFREAISQTFPWVLTCCQFDDKETEDDKDAETEIPAGESSDAAPSA
DAAQMKEMMAMMQKMQQQQAAYPPQGYAPPPQGYPPQGYPPQGYPPQGYPPQGYPPPPQGAPPQGAPPAAPPQGVDNQAYQA
1|32|57|2|72|97|3|106|131|4|152|172|5|197|221|6|262|286|7|295|317

4pxz_A|466|7|2.5|X-RAY|0.52|310.59|38.82
DYKDDDDGAPQAVDNLTSAPGNTSLCTRDYKITQVLFPLLYTVLFFVGLITNGLAMRIFFQIRSKSNFIIFLKNTVISDLLMILTFPFKILSDAKLGTGPLRTFVCQVTSVIFYFTMYISIS
FLGLITIDRYQKTTRPFKTSNPKNLLGAKILSVVIWAFMFLLSLPNMILTNRQPRDKNVKKCSFLKSEFGLVWHEIVNYICQVIFWINFLIVIVCYTLITKELYRSYVRTADLEDNWETLND
NLKVIEKADNAAQVKDALTKMRAAALDAQKATPPKLEDKSPDSPEMKDFRHGFDILVGQIDDALKLANEGKVKEAQAAAEQLKTTRNAYIQKYLRGVGKVPRKKVNVKVFIIIAVFFICFVP
FHFARIPYTLSQTRDVFDCTAENTLFYVKESTLWLTSLNACLNPFIYFFLCKSFRNSLISMLKCPNSATSLSQDNRKKEQDGGDPNEETPMGRPLEVLFQ**********
****
****
****
****
****
1|22|48|2|61|85|3|91|120|4|140|163|5|182|210|6|237|264|7|274|297

4lp8_A|301|2|2.46|X-RAY|0.72|134.16|11.05
MTGGMKPPARKPRILNSDGSSNITRLGLEKRGWLDDHYHDLLTVSWPVFITLITGLYLVTNALFALAYLACGDVIENARPGSFTDAFFFSVQTMATIGYGKLIPIGPLANTLVTLEALCGML
GLAVAARLIYARFTRPTAGVLFSSRMVISDFEGKPTLMMRLANLRIEQIIEADVHLVLVRSEISQEGMVFRRFHDLTLTRSRLPIFSLSWTVMHPIDHHSPIYGETDETLRNSHSEFLVLFT
GHHEAFAQNVHARHAYSCDEIIWGGHFVDVFTTLPDGRRALDLGKFHEIAQHHHHHH
1|46|71|2|107|131

4uc3_B|155|5|2.5|X-RAY|0.57|94.03|19.06
**MDWALFLTFLAACGAPATTGALLKPDEWYDNLNKPWWNPPRWVFPLAWTSLYFLMSLAAMRVAQLEGSGQALAFYAAQLAFNTLWTPVFFGMKRMATALAVVMVMWLFVAATMWAFFQLD
TWAGVLFVPYLIWATAATGLNFEAMRLNWNRPEAR
1|5|24|2|43|65|3|71|90|4|98|119|5|124|144

1jb0_F|164|1|2.5|X-RAY|0.74|67.57|4.94
MRRFLALLLVLTLWLGFTPLASADVAGLVPCKDSPAFQKRAAAAVNTTADPASGQKRFERYSQALCGEDGLPHLVVDGRLSRAGDFLIPSVLFLYIAGWIGWVGRAYLIAVRNSGEANEKEI
IIDVPLAIKCMLTGFAWPLAALKELASGELTAKDNEITVSPR
1|60|84

3ayf_A|800|14|2.5|X-RAY|0.52|538.3|81.82
MEVNRTVSPNIQTGRKTTNSFLKSILIFTILISSTVLLVGGYWIFKEMAPRPKEVRSESGEVLMTKETIIGGQAVFQKYGLMDYGTVLGHGSYMGPDYTAEALKVYTEGMQDYKAKERYNKP
FADLTDDEKSIIREQVIKEMRKNRYNPVTDVLVLTDAQVYGLEKVRDYYRDVFTNGDGWGLKKGLIKESDMPKANRAWVADSDQIQQIADFFFWTAWLSSTLRIGDEITYTNNWPYYEDAGN
TMSFSAVWWSGASVTILILFIGIILYVFYRYQLSMQEAYAEGKFPVIDLRRQPLTPSQVKAGKYFVVVSALFFVQTMFGALLAHYYTEPDSFFGINWIYDILPFNIAKGYHLQLAIFWIATA
WLGMGIFIAPLVGGQEPKKQGLLVDLLFWALVVLVGGSMIGQWLGVNGYLGNEWFLLGHQGWEYIELGRIWQIILVVGMLLWLFIVFRGVKRGLKRESDKGGLIHLLFYSAIAVPFFYIFAF
FIQPDTNFTMADFWRWWIIHLWVEGIFEVFAVVVIGFLLVQLRLVTKKSTVRALYFQFTILLGSGVIGIGHHYYYNGSPEVWIALGAVFSALEVIPLTLLILEAYEQYKMMRDGGANFPYKA
TFWFLISTAIWNLVGAGVFGFLINLPAVSYFEHGQFLTPAHGHAAMMGVYGMFAIAVLLYSLRNIVKPEAWNDKWLKFSCWMLNIGLAGMVVITLLPVGILQMKEAFIHGYWASRSPSFLQQ
DVVQNLLLVRAVPDTIFLIGVVALLVFAIKALFHLRKPTHGEGEELPVANHWMKDRLKNSLEHHHHHH
1|21|44|2|251|272|3|304|331|4|351|372|5|387|410|6|436|453|7|472|490|8|498|522|9|542|561|10|569|592|11|614|639|12|646|669|1
3|685|710|14|736|762

4tq4_A|303|9|2.5|X-RAY|0.53|199.04|40.06
MDSSLANINQIDVPSKYLRLLRPVAWLCFLLPYAVGFGFGITPNASLQHAVLGLLSFAFWMAFSFTINALYDRDVDRLHDGRVKDLNLSMQPLVTGEISVREAWLYCIAFLALSLATAAAIN
EKFFLAMLGANIIGYVYSAPPRFKAWPVMDVICNALAAVLAFYAGLSIGGAEVPIAIYPAAFFLAATFYIPTAVSDYEFDKKAGLKNTPVFFGPERALKSLYPLSAITVILWAYVFLMAERI
EIKVISPLIIAYTLIYTFIINSRWDGEKLNVSPNLILTPFGIISALFIAYGFAVISVLG
1|20|41|2|46|68|3|102|124|4|125|140|5|150|171|6|177|194|7|221|240|8|245|264|9|278|299

1jb0_A|755|11|2.5|X-RAY|0.57|466.29|65.29
MTISPPEREPKVRVVVDNDPVPTSFEKWAKPGHFDRTLARGPQTTTWIWNLHALAHDFDTHTSDLEDISRKIFSAHFGHLAVVFIWLSGMYFHGAKFSNYEAWLADPTGIKPSAQVVWPIVG
QGILNGDVGGGFHGIQITSGLFQLWRASGITNEFQLYCTAIGGLVMAGLMLFAGWFHYHKRAPKLEWFQNVESMLNHHLAGLLGLGSLAWAGHQIHVSLPINKLLDAGVAAKDIPLPHEFIL
NPSLMAELYPKVDWGFFSGVIPFFTFNWAAYSDFLTFNGGLNPVTGGLWLSDTAHHHLAIAVLFIIAGHMYRTNWGIGHSLKEILEAHKGPFTGAGHKGLYEVLTTSWHAQLAINLAMMGSL
SIIVAQHMYAMPPYPYLATDYPTQLSLFTHHMWIGGFLVVGGAAHGAIFMVRDYDPAMNQNNVLDRVLRHRDAIISHLNWVCIFLGFHSFGLYVHNDTMRAFGRPQDMFSDTGIQLQPVFAQ
WVQNLHTLAPGGTAPNAAATASVAFGGDVVAVGGKVAMMPIVLGTADFMVHHIHAFTIHVTVLILLKGVLFARSSRLIPDKANLGFRFPCDGPGRGGTCQVSGWDHVFLGLFWMYNCISVVI
FHFSWKMQSDVWGTVAPDGTVSHITGGNFAQSAITINGWLRDFLWAQASQVIGSYGSALSAYGLLFLGAHFIWAFSLMFLFSGRGYWQELIESIVWAHNKLKVAPAIQPRALSIIQGRAVGV
AHYLLGGIATTWAFFLARIISVG
1|72|93|2|159|179|3|193|216|4|298|314|5|354|374|6|392|413|7|439|464|8|536|557|9|591|613|10|674|691|11|725|745

2bhw_A|232|3|2.5|X-RAY|0.59|135.69|14.68
RKSATTKKVASSGSPWYGPDRVKYLGPFSGESPSYLTGEFPGDYGWDTAGLSADPETFSKNRELEVIHSRWAMLGALGSVFPELLSRNGVKFGEAVWFKAGSQIFSEGGLDYLGNPSLVHAQ
SILAIWATQVILMGAVEGYRIAGGPLGEVVDPLYPGGSFDPLGLADDPEAFAELKVKELKNGRLAMFSMFGFFVQAIVTGKGPLENLADHLADPVNNNAWSYATNFVPGK
1|63|85|2|124|143|3|178|200

4phz_A|420|2|2.59|X-RAY|0.82|133.8|11.88
MKKLVKLAAFGAAAAVAATLGAVAPASAHGEKSQQAFLRMRTLNWYDVQWSKTTVNVNEEMVLSGKVHVFSAWPQAVANPRVSFLNAGEPGPVLVRTAQFIGEQFAPRSVSLEIGKDYAFSI
NLRGRRAGRWHVHAQINVEGGGPIIGPGQWIEIKGDMKDFTDPVTLLDGSTVDLEHYGISRVYAWHLPWMAVGAAWIFFWFVRKGIITSYIRVAEGKADDVIGDDDRRIGAIVLALTILATI
VGYAVTNSTFPRTIPLQAGLQKPLTPIETEGTVGVGKENVTTELNGGVYKVPGRELTINVKVKNNTSQPLRLGEYTAAGLRFLNPDVFTTKPDFPDYLLADRGLSVDATPIAPGEAKEIVVK
IQDARWDIERLSDLAYDTDSQIGGLLFFFSPDGKRYASEIGGPVIPKFVAGDMP
1|184|203|2|228|249

4phz_B|252|6|2.59|X-RAY|0.5|171.48|28.25
MSQSKSGGAVGPFNSVAEAAGCVQTVDWMLLVLLFFAVLGGYHVHFMLTAGDWDFWVDWKDRRMWPTVVPILGVTFCAASQAFWWVNFRLPFGAVFAALGLLIGEWINRYVNFWGWTYFPIS
LVFPSALIVPAIWLDVILLLSGSYVITAVVGSLGWGLLFYPNNWPAIAAFHQATEQHGQLMTLADLIGFHFVRTSMPEYIRMVERGTLRTFGKDVVPVAAFFSGFVSMMVYFLWWFMGRWYS
TTKVIDTI
1|28|51|2|65|85|3|93|113|4|130|145|5|146|168|6|219|244

1nek_B|238|3|2.6|X-RAY|0.57|145.42|14.62
MRLEFSIYRYNPDVDDAPRMQDYTLEADEGRDMMLLDALIQLKEKDPSLSFRRSCREGVCGSDGLNMNGKNGLACITPISALNQPGKKIVIRPLPGLPVIRDLVVDMGQFYAQYEKIKPYLL
NNGQNPPAREHLQMPEQREKLDGLYECILCACCSTSCPSFWWNPDKFIGPAGLLAAYRFLIDSRDTETDSRLDGLSDAFSVFRCHSIMNCVSVCPKGLNPTRAIGHIKSMLLQRNA
1|17|39|2|55|80|3|88|113

4zr1_A|298|4|2.6|X-RAY|0.6|174.26|21.51
**********MARLLTNKNHKVEVQLSADGTEFDSTTF
VKELPAEEKLSIATDYSNDYKKHKFLDLNRPLLMQILRSDFKKDFYVDQIHRPRHYGKGSAPLFGNFLEPLTKTAWWVVPVAWLPVVVYHMGVALKNMNQLFACFLFCVGVFVWTLIEYGLH
RFLFHFDDWLPESNIAFATHFLLHGCHHYLPMDKYRLVMPPTLFVILCAPFYKLVFALLPLYWAYAGFAGGLFGYVCYDECHFFLHHSKLPPFMRKLKKYHLEHHYKNYQLGFGVTSWFWDE
VFGTYLGPDAPLSKMKYESGLEVLFQ
1|199|214|2|222|248|3|285|303|4|307|328

4xes_A|541|7|2.6|X-RAY|0.57|330.26|37.98
**********DYKDDDDATSTSESDTAGPNSDLDVNTDIYSKVLVTAIYLALFVVGTVGNSVTLFTLARKKSLQSLQSTVHYHLGSLALSDLLILLLAMP
VELYNFIWVHHPWAFGDAGCRGYYFLRDACTYATALNVASLSVARYLAICHPFKAKTLMSRSRTKKFISAIWLASALLAIPMLFTMGLQNRSADGTHPGGLVCTPIVDTATVKVVIQVNTFM
SFLFPMLVISILNTVIANKLTVMVHQAAEQGRVCTVGTHNGLEHSTFNMTIEPGRVQALRHGVLVLRAVVIAFVVCWLPYHVRRLMFCYISDEQWTTFLFDFYHYFYMLTNALFYASSAINP
ILYNLVSANFRQVFLSTLACLCPGWRHRRKAHHHHHHHHHHGGNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGILR
NAKLKPVYDSLDAVRRAALINMVFQMGETGVAGFTNSLRMLNNKRWDEAAVNLAKSRWYNQTPNRAKRVITTFRTGTWDAYGSGS**********
****
****
****
****
****
1|61|86|2|105|129|3|140|165|4|187|207|5|233|258|6|310|330|7|349|369

4gx0_A|565|2|2.6|X-RAY|0.85|159.06|12.51
***MQRGSAYFLRGRARQNLKVLLLYCAFLLVMLLAYASIFRYLMWHLEGRAYSFMAGIYWTITVMTTLGFGDITFESDAGYLFASIVTVSGVIFLDIILPFGFVSMFLAPWIERRLRYHPT
IELPDDTRGHILIFGIDPITRTLIRKLESRNHLFVVVTDNYDQALHLEEQEGFKVVYGSPTDAHVLAGLRVAAARSIIANLSDPDNANLCLTVRSLCQTPIIAVVKEPVHGELLRLAGANQV
VPLTRILGRYLGIRATTCGALAHILDSFGNLQIAELPVHGTPFAGKTIGESGIRQRTGLSIIGVWERGSLTTPQRETVLTEQSLLVLAGTKSQLAALEYLIGEAPEDELIFIIGHGRIGCAA
AAFLDRKPVPFILIDRQESPVCNDHVVVYGDATVGQTLRQAGIDRASGIIVTTNDDSTNIFLTLACRHLHSHIRIVARANGEENVDQLYAAGADFVVSNASVGANILGNLLEHKESAFLSEG
MAVFRRPLPPAMAGKTIAETRLRPLTGCSIVAIEAPDRADILISPPPETILAEGARLILIGTSEQEKTFDQTIAARLVPR
1|20|41|2|81|105

3jqo_A|227|1|2.6|X-RAY|0.85|60.56|5.35
****
**********SNSPGAQPQDNETSEGSSALAKNLTPARLKASRAGVMANPSLTVPKGKMIPCGTGTELDTTVPGQVSCRVSQDVYSADGLVRLID
KGSWVDGQITGGIKDGQARVFVLWERIRNDQDGTIVNIDSAGTNSLGSAGIPGQVDAHMWERLRGAIMISLFSDTLTALVNQTQSNNIQYNSTENSGGQLASEALRSYMSIPPTLYDQQGDA
VSIFVARDLDFSGVYTLADN
1|307|324

4av3_A|735|16|2.6|X-RAY|0.5|504.17|88.87
MRGSHHHHHHYVAALFFLIPLVALGFAAANFAAVVRKPEGTERMKEISSYIRSGADSFLAHETKAIFKVAIVIAILLMIFTTWQTGVAFLLGAVMSASAGIVGMKMATRANVRVAEAARTTK
KIGPALKVAYQGGSVMGLSVGGFALLGLVLVYLIFGKWMGQVDNLNIYTNWLGINFVPFAMTVSGYALGCSIIAMFDRVGGGVYTKAADMAADLVGKTELNLPEDDPRNPATIADNVGDNVG
DVAGLGADLLESFVGAIVSSIILASYMFPIYVQKIGENLVHQVPKETIQALISYPIFFALVGLGCSMLGILYVIVKKPSDNPQRELNISLWTSALLTVVLTAFLTYFYLKDLQGLDVLGFRF
GAISPWFSAIIGIFSGILIGFWAEYYTSYRYKPTQFLGKSSIEGTGMVISNGLSLGMKSVFPPTLTLVLGILFADYFAGLYGVAIAALGMLSFVATSVSVDSYGPIADNAGGISEMCELDPE
VRKITDHLDAVGNTTAAIGKGFAIGSAIFAALSLFASYMFSQISPSDIGKPPSLVLLLNMLDARVIAGALLGAAITYYFSGYLISAVTKAAMKMVDEIRRQAREIPGLLEGKAKPDYNRCIE
ITSDNALKQMGYPAFIAILTPLVTGFLLGAEFVGGVLIGTVLSGAMLAILTANSGGAWDNAKKYLEAGNLEGYGKGSEPHKALVIGDTVGDPLKDTVGPSLDILIKIMSVVSVIAVSIFKHV
HLF
1|4|20|2|54|74|3|75|93|4|128|152|5|171|192|6|239|261|7|286|307|8|323|344|9|359|379|10|416|436|11|437|456|12|498|521|13|539
|559|14|611|632|15|633|650|16|699|722

1p49_A|562|2|2.6|X-RAY|0.86|151.54|12.51
**********HAASRPNIILVMADDLGIGDPGCYGNKTIRTPNIDRLASGGVKLTQHLAASPLATPSRAAFMTGRYPVRSGMASWSRTGVFLFTASSGGLPTDEITFAKLL
KDQGYSTALIGKWHLGMSCHSKTDFCHHPLHHGFNYFYGISLTNLRDCKPGEGSVFTTGFKRLVFLPLQIVGVTLLTLAALNCLGLLHVPLGVFFSLLFLAALILTLFLGFLHYFRPLNCFM
MRNYEIIQQPMSYDNLTQRLTVEAAQFIQRNTETPFLLVLSYLHVHTALFSSKDFAGKSQHGVYGDAVEEMDWSVGQILNLLDELRLANDTLIYFTSDQGAHVEEVSSKGEIHGGSNGIYKG
GKANNWEGGIRVPGILRWPRVIQAGQKIDEPTSNMDIFPTVAKLAGAPLPEDRIIDGRDLMPLLEGKSQRSDHEFLFHYCNAYLNAVRWHPQNSTSIWKAFFFTPNFNPVGSNGCFATHVCF
CFGSYVTHHDPPLLFDISKDPRERNPLTPASEPRFYEILKVMQEAADRHTQTLPEVPDQFSWNNFLWKPWLQLCCPSTGLSCQCDREKQDKRLSR
1|183|206|2|213|232

1nek_A|588|3|2.6|X-RAY|0.78|221.43|18.04
MKLPVREFDAVVIGAGGAGMRAALQISQSGQTCALLSKVFPTRSHTVSAQGGITVALGNTHEDNWEWHMYDTVKGSDYIGDQDAIEYMCKTGPEAILELEHMGLPFSRLDDGRIYQRPFGGQ
SKNFGGEQAARTAAAADRTGHALLHTLYQQNLKNHTTIFSEWYALDLVKNQDGAVVGCTALCIETGEVVYFKARATVLATGGAGRIYQSTTNAHINTGDGVGMAIRAGVPVQDMEMWQFHPT
GIAGAGVLVTEGCRGEGGYLLNKHGERFMERYAPNAKDLAGRDVVARSIMIEIREGRGCDGPWGPHAKLKLDHLGKEVLESRLPGILELSRTFAHVDPVKEPIPVIPTCHYMMGGIPTKVTG
QALTVNEKGEDVVVPGLFAVGEIACVSVHGANRLGGNSLLDLVVFGRAAGLHLQESIAEQGALRDASESDVEASLDRLNRWNNNRNGEDPVAIRKALQECMQHNFSVFREGDAMAKGLEQLK
VIRERLKNARLDDTSSEFNTQRVECLELDNLMETAYATAVSANFRTESRGAHSRFDFPDRDDENWLCHSLYLPESESMTRRSVNMEPKLRPAFPPKIRTY
1|17|39|2|55|80|3|88|113

5tin_A|362|4|2.61|X-RAY|0.63|197.75|21.8
ARSRSAPMSPSDFLDKLMGRTSGYDARIRPNFKGPPVQVTCNIFINSFGSIAETTMDYRVNIFLRQKWNDPRLAYSEYPDDSLDLDPSMLDSIWKPDLFFANEKGANFHEVTTDNKLLRIFK
NGNVLYSIRLTLTLSCPMDLKNFPMDVQTCIMQLESFGYTMNDLIFEWQDEAPVQVAEGLTLPQFLLKEEKDLRYCTKHYNTGKFTCIEVRFHLERQMGYYLIQMYIPSLLIVILSWVSFWI
NMDAAPARVALGITTVLTMTTQSSGSRASLPKVSYVKAIDIWMAVCLLFVFSALLEYAAVNFVSRAGTKVFIDRAKKIDTISRACFPLAFLIFNIFYWVIYKILRHEDIHWSHPQFEK
1|224|244|2|249|271|3|283|304|4|322|343

4ymk_A|339|4|2.61|X-RAY|0.63|186.81|21.51
**********MSGNEREKVKTVPLHLEEDIRPEMKEDIHDPTYQDEEGPPPKLEYVWRNIILMVLLHLGGLYGIILVPSCKLYTCLFGIFYYMTSALGITAGAHRLWSHR
TYKARLPLRIFLIIANTMAFQNDVYEWARDHRAHHKFSETHADPHNSRRGFFFSHVGWLLVRKHPAVKEKGGKLDMSDLKAEKLVMFQRRYYKPGLLLMCFILPTLVPWYCWGETFVNSLFV
STFLRYTLVLNATWLVNSAAHLYGYRPYDKNIQSRENILVSLGAVGEGFHNYHHTFPFDYSASEYRWHINFTTFFIDCMAALGLAYDRKKVSKATVLARIKRTGDGSHKSSENLYFQ
1|69|90|2|94|112|3|213|233|4|238|259

4p02_A|803|8|2.65|X-RAY|0.67|411.15|49.18
MGTVRAKARSPLRVVPVLLFLLWVALLVPFGLLAAAPVAPSAQGLIALSAVVLVALLKPFADKMVPRFLLLSAASMLVMRYWFWRLFETLPPPALDASFLFALLLFAVETFSISIFFLNGFL
SADPTDRPFPRPLQPEELPTVDILVPSYNEPADMLSVTLAAAKNMIYPARLRTVVLCDDGGTDQRCMSPDPELAQKAQERRRELQQLCRELGVVYSTRERNEHAKAGNMSAALERLKGELVV
VFDADHVPSRDFLARTVGYFVEDPDLFLVQTPHFFINPDPIQRNLALGDRCPPENEMFYGKIHRGLDRWGGAFFCGSAAVLRRRALDEAGGFAGETITEDAETALEIHSRGWKSLYIDRAMI
AGLQPETFASFIQQRGRWATGMMQMLLLKNPLFRRGLGIAQRLCYLNSMSFWFFPLVRMMFLVAPLIYLFFGIEIFVATFEEVLAYMPGYLAVSFLVQNALFARQRWPLVSEVYEVAQAPYL
ARAIVTTLLRPRSARFAVTAKDETLSENYISPIYRPLLFTFLLCLSGVLATLVRWVAFPGDRSVLLVVGGWAVLNVLLVGFALRAVAEKQQRRAAPRVQMEVPAEAQIPAFGNRSLTATVLD
ASTSGVRLLVRLPGVGDPHPALEAGGLIQFQPKFPDAPQLERMVRGRIRSARREGGTVMVGVIFEAGQPIAVRETVAYLIFGESAHWRTMREATMRPIGLLHGMARILWMAAASLPKTARDF
MDEPARRRRRHEEPKEKQAHLLAFGTDFSTEPDWAGELLDPTAQVSARPNTVAWGSNHHHHHHKLHHHHHH
1|15|34|2|41|60|3|64|86|4|96|120|5|418|435|6|447|466|7|522|543|8|549|570

4ntf_A|156|4|2.65|X-RAY|0.51|104.56|16.44
MHHHHHHKDEVALLATVTLVGVLLQAYFSLQVISARRAFHVSPPLTSGPPEFERVFRAQVNCSEYFPLFLATLWVAGIFFHEGAAALCGLFYLFARLRYFQGYARSAQLRLTPLYASARALW
LLVAMAALGLLVHFLPGTLRTALFRWLQMLLPMA
1|3|22|2|56|75|3|76|92|4|107|135

4umw_A|732|8|2.7|X-RAY|0.67|369.27|50.29
MSTPDNHGKKAPQFAAFKPLTTVQNANDCCCDGACSSSPTLSENVSGTRYSWKVSGMDCAACARKVENAVRQLAGVNQVQVLFATEKLVVDADNDIRAQVESAVQKAGYSLRDEQAADEPQA
SRLKENLPLITLIVMMAISWGLEQFNHPFGQLAFIATTLVGLYPIARQALRLIKSGSYFAIETLMSVAAIGALFIGATAEAAMVLLLFLIGERLEGWAASRARQGVSALMALKPETATRLRN
GEREEVAINSLRPGDVIEVAAGGRLPADGKLLSPFASFDESALTGESIPVERATGDKVPAGATSVDRLVTLEVLSEPGASAIDRILKLIEEAEERRAPIERFIDRFSRIYTPAIMAVALLVT
LVPPLLFAASWQEWIYKGLTLLLIGCPCALVISTPAAITSGLAAAARRGALIKGGAALEQLGRVTQVAFDKTGTLTVGKPRVTAIHPATGISESELLTLAAAVEQGATHPLAQAIVREAQVA
ELAIPTAESQRALVGSGIEAQVNGERVLICAAGKHPADAFAGLINELESAGQTVVLVVRNDDVLGIIALQDTLRADAATAISELNALGVKGVILTGDNPRAAAAIAGELGLEFKAGLLPEDK
VKAVTKLNQHAPLAMVGDGINDAPAMKAAAIGIAMGSGTDVALETADAALTHNHLRGLVQMIELARATHANIRQNITIALGLKGIFLVTTLLGMTGLWLAVLADTGATVLVTANALRLLRRR
1|129|148|2|150|169|3|181|196|4|200|213|5|353|373|6|377|399|7|684|702|8|707|724

4quv_A|427|10|2.74|X-RAY|0.5|291.12|49.22
MSEQESRDNAAVDAVRQKYGFGFSWLVLMIALPPLVYYLWICVTYYQGELVFTSDAAAWRRFWSHVAPPTWHAAGLYAAWFLGQAALQVWAPGPTVQGMKLPDGSRLDYRMNGIFSFLFTLA
VVFGLVTMGWLDATVLYDQLGPLLTVVNIFTFVFAGFLYFWGLNGKQWERPTGRPFYDYFMGTALNPRIGSLDLKLFCEARPGMIFWLLMNLSMAAKQYELHGTVTVPMLLVVGFQSFYLID
YFIHEEAVLTTWDIKHEKFGWMLCWGDLVWLPFTYTLQAQYLVHHTHDLPVWGIIAIVALNLAGYAIFRGANIQKHHFRRDPNRIVWGKPAKYIKTKQGSLLLTSGWWGIARHMNYFGDLMI
ALSWCLPAAFGSPIPYFHIVYFTILLLHREKRDDAMCLAKYGEDWLQYRKKVPWRIVPKIY
1|24|43|2|71|90|3|113|130|4|143|162|5|196|217|6|229|248|7|265|286|8|295|314|9|358|374|10|380|395

5fgn_A|550|5|2.75|X-RAY|0.7|260.63|30.56
MIKPNLRPKLGSSALIAFLSLYSSLVLNYAFFAKVVELHPFNGTGADIFLYTMPVVLFFLSNFVFHVIALPFVHKVLIPLILVISAAVSYQEIFFNIYFNKSMLNNVLQTTAAESARLITPG
YVLWIVCLGVLPALAYIAVKVKYRVWYKEFLTRLVLAAVSFLCALGIAMLQYQDYASFFRNNKSVTHLIVPSNFIGAGVSKYKDWKRSNIPYTQLDMAVVQNRPAGSLRRFVVLVVGETTRA
ANWGLNGYSRQTTPLLAARGDEIVNFPQVRSCGTSTAHSLPCMFSTFDRTDYDEIKAEHQDNLLDIVQRAGVEVTWLENDSGCKGVCGKVPNTDVTSLNLPEYCRNGECLDNILLTKFDEVL
NKNDKDAVLILHTIGSHGPTYYERYTEAERKFTPTCDTNEINKCTRATLVNTYDNTVLYVDQFIDKVIRKLENRDDLESVVHYVSDHGESLGENGMYLHAAPYAIAPSGQTHIPMVMWFSKA
FRQHGGIDFQCLKQKAAENEYSHDHYFSTVLGLMDISNSQTYRKEMDILAACRRPRHHHHHH
1|14|29|2|50|70|3|73|90|4|119|140|5|148|172

4yuu_D1|351|5|2.77|X-RAY|0.55|222.33|26.58
*MTIAIGREQERGWFDLLDDWLKRDRFVFIGWSGILLFPCAYLALGAWFTGTTFVSSWYTHGLASSYLEGCNFLTAAVSSPANSMGHSLLFLWGPEAQGDFTRWCQIGGLWTFTALHGSFGL
IGFCLRQFEIARLVGLRPYNAIAFSGPIAVFVSVFLLYPLGQASWFFAPSFGVAAIFRFLLFLQGFHNWTLNPFHMMGVAGILGGALLCAIHGATVENTLFEDGEASDTFRAFTPTQSEETY
SMVTANRFWSQIFGVAFANKRWLHFFLLFVPVTGLWVSSIGIVGLALNLRAYDFVSQEIRAAEDPEFETFYTKNILLNEGIRAWMAAQDQPHENFVFPEEVLPRGNAL
1|32|53|2|109|132|3|141|162|4|191|217|5|266|290

4yuu_M1|108|1|2.77|X-RAY|0.68|52.24|4.47
MLGFLSNHLVLCRSRLPLQNNKFTSAKRVGGATYVARPCMLLNILPVQEGVKHLTHIAPMYLSSESGGMVVQEGGYLAVLLGVLFPVAFLIILYIQSEARNAGMREAA
1|6|27

5x3x_M|222|3|2.79|X-RAY|0.64|119.25|15.42
MHIMEGYLPVTHAIGWSLAAAPFVVAGALKIRKIVAERPEARMTLAAAGAFAFVLSALKIPSVTGSCSHPTGTGLGAVVFGPSVMAVLGVIVLLFQALLLAHGGLTTLGANAFSMAIVGPWV
AFGVYKLAGKAGASMAVAVFLAAFLGDLATYVTTSLQLALAYPDPASGFLGAALKFGSVFALTQIPLAIAEGFLTVIVVDALAGKVADEDKLRILAGEAR
1|11|23|2|146|164|3|171|191

1yew_C|289|5|2.8|X-RAY|0.58|173.82|25.92
MHETKQGGEKRFTGAICRCSHRYNSMEVKMAATTIGGAAAAEAPLLDKKWLTFALAIYTVFYLWVRWYEGVYGWSAGLDSFAPEFETYWMNFLYTEIVLEIVTASILWGYLWKTRDRNLAAL
TPREELRRNFTHLVWLVAYAWAIYWGASYFTEQDGTWHQTIVRDTDFTPSHIIEFYLSYPIYIITGFAAFIYAKTRLPFFAKGISLPYLVLVVGPFMILPNVGLNEWGHTFWFMEELFVAPL
HYGFVIFGWLALAVMGTLTQTFYSFAQGGLGQSLCEAVDEGLIAK
1|49|69|2|91|109|3|136|156|4|174|193|5|253|258

4xk8_L|153|3|2.8|X-RAY|0.5|102.76|13.33
***YQVIQPINGDPFIGSLETPVTSSPLIAWYLSNLPAYRTAVSPLLRGIEVGLAHGYLLVGPFVKAGPLRNTEIAGQAGSLAAGGLVVILSLCLTIYGISSFNEGAPSTAPSLTLTGRKKE
PDQLQTADGWAKFTGGFFFGGISGVIWAYFLLYV
1|45|69|2|77|98|3|132|154

4kt0_J|40|1|2.8|X-RAY|0.56|23.21|2.64
MDGLKSFLSTAPVMIMALLTFTAGILIEFNRFYPDLLFHP
1|7|33

4xk8_H|90|1|2.8|X-RAY|0.67|44.11|4.28
**SVYFDLEDLGNTTGQWDSYGSDAPSPYNPLQSKLFETFAAPFTKRGLLLKFLILGGGSTLAYLSATAS
GDILPITRGPQQPPKLGPRGKI
1|101|119

4dw1_A|340|2|2.8|X-RAY|0.79|122.44|11.41
**********GSSKKVGTLNRFTQALVIAYVIGYVCVYNKGYQDTDTVLSSVTTKVKGIALTKTSELGERIWDVADYIIPPQEDGSFFVLTNMIITTNQTQSKCAEN
PTPASTCTSHRDCKRGFNDARGDGVRTGRCVSYSASVKTCEVLSWCPLEKIVDPPNPPLLADAERFTVLIKNNIRYPKFNFNKRNILPNINSSYLTHCVFSRKTDPDCPIFRLGDIVGEAEE
DFQIMAVRGGVMGVQIRWDCDLDMPQSWCVPRYTFRRLDNKDPDNNVAPGYNFRFAKYYKNSDGTETRTLIKGYGIRFDVMVFGQAGKFNIIPTLLNIGAGLALLGLVNVICDWIVLTFMK
1|38|56|2|335|356

4kt0_K|128|2|2.8|X-RAY|0.55|79.76|8.9
MTAIARERGRSKRGIALHRLEYLKENQVFEIIFGEPLSMFNTALLLAQASPTTAGWSLSVGIIMCLCNVFAFVIGYFAIQKTGKGKDLALPQLASKKTFGLPELLATMSFGHILGAGMVLGL
ASSGIL
1|56|75|2|102|125

4rdq_A|409|4|2.85|X-RAY|0.69|196.29|23.24
*TVTYTNRVADARLGTFSQLLLQWKGSIYKLLYSEFLIFISLYFAISLVYRLILSESQRLMFEKLALYCNSYAELIPVSFVLGFYVSLVVSRWWAQYESIPWPDRIMNLVSCNVDGEDEYGR
LLRRTLMRYSNLCSVLILRSVSTAVYKRFPSMEHVVRAGLMTPEEHKKFESLNSPHNKFWIPCVWFSNLAVKARNEGRIRDSVLLQGILNELNTLRSQCGRLYGYDWISIPLVYTQVVTVAV
YSFFLACLIGRQFLDPEKAYPGHELDLFVPVFTFLQFFFYAGWLKVAEQLINPFGEDDDDFETNWLIDRNLQVSLMAVDEMHQDLPILEKDLYWNEPDPQPPYTAATAEYKRPSFLGSTFDI
SMQKEEMEFQPLEQIKENEEANHSTPLLGHLGRLLGVQSEGEEF
1|31|51|2|61|82|3|238|255|4|272|288

4xyd_B|150|1|2.85|X-RAY|0.75|60.14|4.86
MRDVLTTSMARNIFYGGSLFFILIFVGLSVHSHRYIVTTSTDAATLTAEVEHGKHLWEIHGCVNCHSILGEGAYFAPELGNVMTRWGVEDDPDAAFEALKGWMDAMPTGIEGRRQMPNFGLN
DEEYRALSDFLLWTNTIRNQDWPPNDAG
1|9|30

5dqq_A|723|12|2.87|X-RAY|0.53|478.66|69.45
**********MGGGGTDRVRRSEAITHGTPFQKAAALVDLAEDGIGLPVEILDQSSFGESARYYFIFTRLDLIWSLNYFALLFLNFFEQPLWCEKNPKPSCKDRDYYYLGELPYLTNAESII
YEVITLAILLVHTFFPISYEGSRIFWTSRLNLVKVACVVILFVDVLVDFLYLSPLAFDFLPFRIAPYVRVIIFILSIRELRDTLVLLSGMLGTYLNILALWMLFLLFASWIAFVMFEDTQQG
LTVFTSYGATLYQMFILFTTSNNPDVWIPAYKSSRWSSVFFVLYVLIGVYFVTNLILAVVYDSFKEQLAKQVSGMDQMKRRMLEKAFGLIDSDKNGEIDKNQCIKLFEQLTNYRTLPKISKE
EFGLIFDELDDTRDFKINKDEFADLCQAIALRFQKEEVPSLFEHFPQIYHSALSQQLRAFVRSPNFGYAISFILIINFIAVVVETTLDIEESSAQKPWQVAEFVFGWIYVLEMALKIYTYGF
ENYWREGANRFDFLVTWVIVIGETATFITPDENTFFSNGEWIRYLLLARMLRLIRLLMNVQRYRAFIATFITLIPSLMPYLGTIFCVLCIYCSIGVQVFGGLVNAGNKKLFETELAEDDYLL
FNFNDYPNGMVTLFNLLVMGNWQVWMESYKDLTGTWWSITYFVSFYVITILLLLNLVVAFVLEAFFTELDLEEEEKCQGQDSQEKRNRRRSAGSKSRSQRVDTLLHHMLGDELSKPECSTSD
T
1|71|91|2|119|139|3|150|172|4|187|204|5|212|234|6|279|308|7|432|453|8|462|484|9|494|517|10|527|548|11|560|583|12|649|676

2a79_B|499|5|2.9|X-RAY|0.7|235.42|30.08
MTVATGDPVDEAAALPGHPQDTYDPEADHECCERVVINISGLRFETQLKTLAQFPETLLGDPKKRMRYFDPLRNEYFFDRNRPSFDAILYYYQSGGRLRRPVNVPLDIFSEEIRFYELGEEA
MEMFREDEGYIKEEERPLPENEFQRQVWLLFEYPESSGPARIIAIVSVMVILISIVSFCLETLPIFRDENEDMHGGGVTFHTYSNSTIGYQQSTSFTDPFFIVETLCIIWFSFEFLVRFFAC
PSKAGFFTNIMNIIDIVAIIPYFITLGTELAEKPEDAQQGQQAMSLAILRVIRLVRVFRIFKLSRHSKGLQILGQTLKASMRELGLLIFFLFIGVILFSSAVYFAEADERDSQFPSIPDAFW
WAVVSMTTVGYGDMVPTTIGGKIVGSLCAIAGVLTIALPVPVIVSNFNYFYHRETEGEEQAQYLQVTSCPKIPSSPDLKKSRSASTISKSDYMEIQEGVNNSNEDFREENLKTANCTLANTN
YVNITKMLTDV
1|222|243|2|294|310|3|325|349|4|364|373|5|385|402

5v56_A|653|7|2.9|X-RAY|0.64|354.55|41.74
**********AVTGPPPPLSHCGRAAPCEPLRYNVCLGSVLPYGATSTLLAGDSDSQEEAHGKLVLWSGLRNAPRCWAVI
QPLLCAVYMPKCENDRVELPSRTLCQATRGPCAIVERERGWPDFLRCTPDRFPEGCTNEVQNIKFNSSGQCMVPLVRTDNPKSWYEDVEGCGIQCQNPLFTEAEHQDMHSYIAAFGAVTGLC
TLFTLATFVADWRNSNRYPAVILFYVNACFFVGSIGWLAQFMDGARREIVCRADGTMRLGEPTSNETLSCVIIFVIVYYALMAGVVWFVVLTYAWHTSFKALGTTYQPLSGKTSYFHLLTWS
LPFVLTVAILAVAQVDGDSVSGICFVGYKNYRYRAGFVLAPIGLVLIVGGYFLIRGVMTLFSIKSNHAKALIVYGSTTGNTEYTAETIARELADAGYEVDSRDAASVEAGGLFEGFDLVLLG
CSTWGDDSIELQDDFIPLFDSLEETGAQGRKVACFGCGDSSWEYFCGAVDAIEEKLKNLGAEIVQDGLRIDGDPRAARDDIVGWAHDVRGAIKINETMLRLGIFGFLAFGFVLITFSCHFYD
FFNQAEWERSFRDYVLCQANVTIGLPTKQPIPDCEIKNRPSLLVEKINLFAMFGTGIAMSTWVWTKATLLIWRRTWCRLTGQSDDHHHHHHHHHH
1|232|254|2|265|285|3|313|336|4|360|378|5|398|420|6|453|475|7|516|537

5mrw_A|557|10|2.9|X-RAY|0.52|371.58|57.58
MAAQGFLLIATFLLVLMVLARPLGSGLARLINDIPLPGTTGVERVLFRALGVSDREMNWKQYLCAILGLNMLGLAVLFFMLLGQHYLPLNPQQLPGLSWDLALNTAVSFVTNTNWRSYSGET
TLSYFSQMAGLTVQNFLSAASGIAVIFALIRAFTRQSMSTLGNAWVDLLRITLWVLVPVALLIALFFIQQGALQNFLPYQAVNTVEGAQQLLPMGPVASQEAIKMLGTNGGGFFNANSSHPF
ENPTALTNFVQMLAIFLIPTALCFAFGEVMGDRRQGRMLLWAMSVIFVICVGVVMWAEVQGNPHLLALGTDSSINMEGKESRFGVLVSSLFAVVTTAASCGAVIAMHDSFTALGGMVPMWLM
QIGEVVFGGVGSGLYGMMLFVLLAVFIAGLMIGRTPEYLGKKIDVREMKLTALAILVTPTLVLMGAALAMMTDAGRSAMLNPGPHGFSEVLYAVSSAANNNGSAFAGLSANSPFWNCLLAFC
MFVGRFGVIIPVMAIAGSLVSKKSQAASSGTLPTHGPLFVGLLIGTVLLVGALTFIPALALGPVAEYLS
1|3|27|2|62|84|3|126|149|4|166|192|5|250|269|6|280|300|7|357|368|8|415|435|9|480|501|10|525|552

5irx_A|636|6|2.95|E-MIC|0.66|330.68|36.6
**********AMGSRLYDRRSIFDAVA
QSNCQELESLLPFLQRSKKRLTDSEFKDPETGKTCLLKAMLNLHNGQNDTIALLLDVARKTDSLKQFVNASYTDSYYKGQTALHIAIERRNMTLVTLLVENGADVQAAANGDFFKKTKGRPG
FYFGELPLSLAACTNQLAIVKFLLQNSWQPADISARDSVGNTVLHALVEVADNTVDNTKFVTSMYNEILILGAKLHPTLKLEEITNRKGLTPLALAASSGKIGVLAYILQREIHEPECRHLS
RKFTEWAYGPVHSSLYDLSCIDTCEKNSVLEVIAYSSSETPNRHDMLLVEPLNRLLQDKWDRFVKRIFYFNFFVYCLYMIIFTAAAYYRPVEGLPPYKLKNTVGDYFRVTGEILSVSGGVYF
FFRGIQYFLQRRPSLKSLFVDSYSEILFFVQSLFMLVSVVLYFSQRKEYVASMVFSLAMGWTNMLYYTRGFQQMGIYAVMIEKMILRDLCRFMFVYLVFLFGFSTAVVTLIEDGKYNSLYST
CLELFKFTIGMGDLEFTENYDFKAVFIILLLAYVILTYILLLNMLIALMGETVNKIAQESKNIWKLQRAITILDTEKSFLKCMRKAFRSGKLLQVGFTPDGKDDYRWCFRVDEVNWTTWNTN
VGIINEDPG**********
1|435|454|2|472|497|3|511|532|4|536|556|5|577|598|6|656|683

5tcx_A|246|4|2.95|X-RAY|0.52|162.02|20.08
RGVEGSTKSIKYLLFVFNFVFWLAGGVILGVALWLRHDPQTTNLLYLELGDKPAPNTFYVGIYILIAVGAVMMFVGFLGCYGAIQESQCLLGTFFTCLVILFACEVAAGIWGFVNKDQIAKD
VKQFYDQALQQAVVDDDANNAKAVVKTFHETLDCCGSSTLTALTTSVLKNNLCPSGSNIISNLFKEDCHQKIDDLFSGKLYLIGIAAIVVAVIMIFEMILSMVLSSGIRNSSVYVPHHHHHH
HH
1|10|37|2|61|79|3|90|113|4|202|227

4m48_A|543|12|2.95|X-RAY|0.5|372.43|64.49
**********MNSISDERETWSGKVDFLLSVIGFAVDLANVWRFPYLCYKNGGGAFLVPYGIMLAVGGIPLFYMELALGQHNRKGAITCWGRLVPLFKGIGYAVVLIAFYVDF
YYNVIIAWSLRFFFASFTNSLPWTSCNNIWNTPNCRPFESQGFQSAASEYFNRYILELNRSEGIHDLGAIKWDMALCLLIVYLICYFSLWKGISTSGKVVWFTALFPYAALLILLIRGLTLP
GSFLGIQYYLTPNFSAIYKAEVWADAATQVFFSLGPGFGVLLAYASYNKYHNNVYKDALLTSFINSATSFIAGFVIFSVLGYMAHTLGVRIEDVATEGPGLVFVVYPAAIATMPASTFWALI
FFMMLATLGLDSSFGGSEAIITALSDEFPKIKRNRELFVAGLFSLYFVVGLASCTQGGFYFFHLLDRYAAGYSILVAVFFEAIAVSWIYGTNRFSEDIRDMIGFPPGRYWQVCWRFVAPIFL
LFITVYLLIGYEPLTYADYVYPSWANALGWCIAGSSVVMIPAVAIFKLLSTPGSLRQRFTILTTPWRDQQLVPR**********
1|34|57|2|62|85|3|105|139|4|237|257|5|260|285|6|309|321|7|342|365|8|404|433|9|444|464|10|480|497|11|517|541|12|555|577

3rko_A|147|3|3.0|X-RAY|0.52|95.9|13.53
MSMSTSTEVIAHHWAFAIFLIVAIGLCCLMLVGGWFLGGRARARSKNVPFESGIDSVGSARLRLSAKFYLVAMFFVIFDVEALYLFAWSTSIRESGWVGFVEAAIFIFVLLAGLVYLVRIGA
LDWTPARSRRERMNPETNSIANRQR
1|17|31|2|73|96|3|97|115

4o9p_C|100|3|2.89|X-RAY|0.52|65.17|11.28
MEFGFWSALYIFVLTAFLGYELITRVPVILHTPLMSGSNFIHGVVVVGAMVVLGHAETGLEKLIGFLGVILGAANAAGGYAVTVRMLEMFERKPGQGGGR
1|6|22|2|36|54|3|60|82

4tkr_A|214|6|3.0|X-RAY|0.52|140.89|26.2
MDYKDDDDKHHHHHHHHHHENLYFQSYVMQNKRLIILLECAIFAAVAMVLSFIPLDIGSSFSISLGMIPMYVIAIRRGFWAAGFAGLLWGLLHFLTGKAYILMPSQAIIEYILAFSFIAFSG
VFSKQVRSNLAANQLKKAIEWAWGTMIIGGVARYFWHYVAGVLFWGAYAFQGWGAQLFSIVMNGASCLGTVLVSGIIISILLKTSPKLFLPK
1|5|25|2|37|50|3|51|68|4|76|97|5|113|139|6|150|176

4pv1_E|32|1|3.0|X-RAY|0.55|18.68|2.48
MILGAVFYIVFIALFFGIAVGIIFAIKSIKLI
1|6|26

4pv1_D|179|1|3.0|X-RAY|0.77|68.1|5.05
MAQFTESMDVPDMGRRQFMNLLAFGTVTGVALGALYPLVKYFIPPSGGAVGGGTTAKDKLGNNVKVSKFLESHNAGDRVLVQGLKGDPTYIVVESKEAIRDYGINAVCTHLGCVVPWNAAEN
KFKCPCHGSQYDETGKVIRGPAPLSLALCHATVQDDNIVLTPWTETDFRTGEKPWWV
1|17|40

4rku_K|72|2|3.0|X-RAY|0.51|47.1|6.33
***DFIGSSTNVIMVASTTLMLFAGRFGLAPSANRKATAGLKLEARDSGLQTGDPAGFTLADTLACGTVGHIIGV
1|49|68|2|97|116

5jsz_D|265|5|3.0|X-RAY|0.57|162.57|26.07
MSKIIIGRYLPGTTFVYRVDPRAKLLTTFYFIIMIFLANNWVSYLVISIFGLAYVFATGLKARVFWDGVKPMIWMIVFTSLLQTFFMAGGKVYWHWWIFTLSSEGLINGLYVFIRFAMIILV
STVMTVTTKPLEIADAMEWMLTPLKLFKVNVGMISLVISIALRFVPTLFDQTVKIMNAQRSRGADFNDGGLVKRAKSVVPMLVPLFIDSLEVALDLSTAMESRGYKGSEGRTRYRILEWSKV
DLIPVAYCLLLTILMITTRKH
1|23|36|2|43|58|3|73|85|4|103|125|5|245|262

4pv1_B|160|3|3.0|X-RAY|0.51|106.03|13.66
MATLKKPDLSDPKLRAKLAKGMGHNYYGEPAWPNDLLYVFPVVIMGTFACIVALSVLDPAMVGEPADPFATPLEILPEWYLYPVFQILRSVPNKLLGVLLMASVPLGLILVPFIENVNKFQN
PFRRPVATTIFLFGTLVTIWLGIGATFPLDKTLTLGLF
1|34|57|2|95|116|3|126|146

1qle_D|43|1|3.0|X-RAY|0.5|27.68|3.09
******TDHKHGEMDIRHQQATFAGFIKGATWVSILSIAVLVFLALANS
1|25|46

2e74_C|289|1|3.0|X-RAY|0.87|72.87|5.59
YPFWAQQTYPPTPREPTGRIVCANCHLAAKPAEVEVPQSVLPDTVFKAVVKIPYDTKLQQVAADGSKVGLNVGAVLMLPEGFKIAPEERIPEELKKEVGDVYFQPYKEGQDNVLLVGPLPGE
QYQEIVFPVLSPNPTTDKNIHFGKYAIHLGANRGRGQIYPTGEKSNNNVFTASATGTITKIAKEEDEYGNVKYQVSIQTDSGKTVVDTIPAGPELIVSEGQAVKAGEALTNNPNVGGFGQDD
TEIVLQDPNRVKWMIAFICLVMLAQLMLILKKKQVEKVQAAEMNF
1|257|277

3h1j_P|380|8|3.0|X-RAY|0.51|257.17|40.61
MAPNIRKSHPLLKMINNSLIDLPAPSNISAWWNFGSLLAVCLMTQILTGLLLAMHYTADTSLAFSSVAHTCRNVQYGWLIRNLHANGASFFFICIFLHIGRGLYYGSYLYKETWNTGVILLL
TLMATAFVGYVLPWGQMSFWGATVITNLFSAIPYIGHTLVEWAWGGFSVDNPTLTRFFALHFLLPFAIAGITIIHLTFLHESGSNNPLGISSDSDKIPFHPYYSFKDILGLTLMLTPFLTLA
LFSPNLLGDPENFTPANPLVTPPHIKPEWYFLFAYAILRSIPNKLGGVLALAASVLILFLIPFLHKSKQRTMTFRPLSQTLFWLLVANLLILTWIGSQPVEHPFIIIGQMASLSYFTILLIL
FPTIGTLENKMLNY
1|30|51|2|83|105|3|111|130|4|181|201|5|225|244|6|289|308|7|321|339|8|350|370

3rko_L|613|18|3.0|X-RAY|0.5|420.95|94.22
MNMLALTIILPLIGFVLLAFSRGRWSENVSAIVGVGSVGLAALVTAFIGVDFFANGEQTYSQPLWTWMSVGDFNIGFNLVLDGLSLTMLSVVTGVGFLIHMYASWYMRGEEGYSRFFAYTNL
FIASMVVLVLADNLLLMYLGWEGVGLCSYLLIGFYYTDPKNGAAAMKAFVVTRVGDVFLAFALFILYNELGTLNFREMVELAPAHFADGNNMLMWATLMLLGGAVGKSAQLPLQTWLADAMA
GPTPVSALIHAATMVTAGVYLIARTHGLFLMTPEVLHLVGIVGAVTLLLAGFAALVQTDIKRVLAYSTMSQIGYMFLALGVQAWDAAIFHLMTHAFFKALLFLASGSVILACHHEQNIFKMG
GLRKSIPLVYLCFLVGGAALSALPLVTAGFFSKDEILAGAMANGHINLMVAGLVGAFMTSLYTFRMIFIVFHGKEQIHAHAVKGVTHSLPLIVLLILSTFVGALIVPPLQGVLPQTTELAHG
SMLTLEITSGVVAVVGILLAAWLWLGKRTLVTSIANSAPGRLLSTWWYNAWGFDWLYDKVFVKPFLGIAWLLKRDPLNSMMNIPAVLSRFAGKGLLLSENGYLRWYVASMSIGAVVVLALLM
VLR
1|3|21|2|30|51|3|84|104|4|114|132|5|138|156|6|164|186|7|220|230|8|236|247|9|258|267|10|282|300|11|305|324|12|329|349|13|37
7|387|14|394|404|15|412|434|16|454|470|17|490|511|18|592|610

5c6p_A|459|12|3.0|X-RAY|0.5|312.66|58.19
****LDRFSFSVFLKEIRLLTALALPMLLAQVAQVGIGFVDTVMAGGAGKEDLAAVALGSSAFATVYITFMGIMAALNPMIAQLYGAGKTGEAGETGRQGIWFGLILGIFGMILMWAAITPF
RNWLTLSDYVEGTMAQYMLFTSLAMPAAMVHRALHAYASSLNRPRLIMLVSFAAFVLNVPLNYIFVYGKFGMPALGGAGCGVATMAVFWFSALALWIYIAKEKFFRPFGLTAKFGKPDWAVF
KQIWKIGAPIGLSYFLEASAFSFIVFLIAPFGEDYVAAQQVGISLSGILYMIPQSVGSAGTVRIGFSLGRREFSRARYISGVSLVSGWVLAVITVLSLVLFRSPLASMYNDDPAVLSIASTV
LLFAGLFQPADFTQCIASYALRGYKVTKVPMFIHAAAFWGCGLLPGYLLAYRFDMGIYGFWTALIASLTIAAVALVWCLEKYSMELVKSHKAVSSGL
1|25|44|2|57|78|3|100|122|4|139|163|5|168|185|6|201|220|7|250|270|8|286|305|9|325|344|10|367|388|11|394|414|12|424|443

1qle_B|252|2|3.0|X-RAY|0.72|112.43|10.7
QDVLGDLPVIGKPVNGGMNFQPASSPLAHDQQWLDHFVLYIITAVTIFVCLLLLICIVRFNRRANPVPARFTHNTPIEVIWTLVPVLILVAIGAFSLPILFRSQEMPNDPDLVIKAIGHQWY
WSYEYPNDGVAFDALMLEKEALADAGYSEDEYLLATDNPVVVPVGKKVLVQVTATDVIHAWTIPAFAVKQDAVPGRIAQLWFSVDQEGVYFGQCSELCGINHAYMPIVVKAVSQEKYEAWLA
GAKEEFAA
1|37|59|2|75|94

2zt9_A|215|4|3.0|X-RAY|0.51|142.93|19.42
MANVYDWFEERLEIQAIAEDVTSKYVPPHVNIFYCLGGITLVCFLIQFATGFAMTFYYKPTVAEAYSSVQYIMNEVNFGWLIRSIHRWSASMMVLMMILHVFRVYLTGGFKKPRELTWVSGV
ILAVITVSFGVTGYSLPWDQVGYWAVKIVSGVPEAIPVVGVLISDLLRGGSSVGQATLTRYYSAHTFVLPWLIAVFMLFHFLMIRKQGISGPL
1|32|53|2|84|107|3|115|134|4|185|207

4ea3_A|434|7|3.01|X-RAY|0.53|286.43|37.8
DYKDDDDGAPADLEDNWETLNDNLKVIEKADNAAQVKDALTKMRAAALDAQKATPPKLEDKSPDSPEMKDFRHGFDILVGQIDDALKLANEGKVKEAQAAAEQLKTTRNAYIQKYLGAFLPL
GLKVTIVGLYLAVCVGGLLGNCLVMYVILRHTKMKTATNIYIFNLALADTLVLLTLPFQGTDILLGFWPFGNALCKTVIAIDYYNMFTSTFTLTAMSVDRYVAICHPIRALDVRTSSKAQAV
NVAIWALASVVGVPVAIMGSAQVEDEEIECLVEIPTPQDYWGPVFAICIFLFSFIVPVLVISVCYSLMIRRLRGVRLLSGSREKDRNLRRITRLVLVVVAVFVGCWTPVQVFVLAQGLGVQP
SSETAVAILRFCTALGYVNSCLNPILYAFLDENFKACFRKFCCASALGRPLEVLFQGPHHHHHHHHHH
1|49|74|2|87|109|3|121|146|4|167|188|5|213|237|6|264|285|7|299|322

3wmm_C|404|1|3.01|X-RAY|0.94|51.86|5.97
MSPAQQLTLPAVIVVASVMLLGCEGPPPGTEQIGYRGVGMENYYNKRQRALSIQANQPVESLPAADSTGPKASEVYQNVQVLKDLSVGEFTRTMVAVTTWVSPKEGCNYCHVPGNWASDDIY
TKVVSRRMFELVRAANSDWKAHVAETGVTCYTCHRGNPVPKYAWVTDPGPKYPSGLKPTGQNYGSKTVAYASLPFDPLTPFLDQANEIRITGNAALAGSNPASLKQAEWTFGLMMNISDSLG
VGCTFCHNTRAFNDWTQSTPKRTTAWYAIRHVRDINQNYIWPLNDVLPASRKGPYGDPLRVSCMTCHQAVNKPLYGAQMAKDYPGLYKTAVTQEALAGSAPASEAAPAAATEAAPEAPAQEV
PAAEAVPAAAEPGAAEAAGSVEPAPVEEVAPAPAAQRL
1|19|30

5doq_A|448|9|3.05|X-RAY|0.5|305.09|46.96
MNGYDPVLLSRILTELTLTVHIIYATIGVGVPLMIAIAQWVGIRKNDMHYILLARRWTRGFVITVAVGVVTGTAIGLQLSLLWPNFMQLAGQVISLPLFMETFAFFFEAIFLGIYLYTWDRF
ENQKKHLLLLIPVAIGSSASAMFITMVNAFMNTPQGFELKNGELVNIDPIVAMFNPAMPTKVAHVLATSYMTSAFVLASIAAWHLWKGNRHIYHRKALHLTMKTAFIFSVASALVGDLSGKF
LAEYQPEKLAAAEWHFETSSHAPLILFGTLEEDNEVKYALEIPYALSILAHNHPAAVVTGLNDIPEDERPPLYIHYLFDVMVTIGVFLMVVAAVYWLGSIFRWKWTAKNWFFGLLVAGGPLA
MIAIEAGWYLAEVGRQPWILRGYMKTAEGATTSAHVDTMLVLFCLLYIVLVIASATVLIRMFRRNPVERELEERANRGEVAP
1|12|37|2|60|79|3|94|113|4|131|151|5|180|202|6|222|242|7|320|342|8|353|378|9|404|425

1q90_R|49|1|3.1|X-RAY|0.5|31.7|3.3
******************************AASSEVPDMNKRNIMNLILAGGAGLPITTLALGYGAFFVPPSSGGGGGG
1|44|66

3ze5_C|130|3|3.1|X-RAY|0.51|86.21|12.3
GHHHHHHELANNTTGFTRIIKAAGYSWKGLRAAWINEAAFRQEGVAVLLAVVIACWLDVDACTRVLLISSVMLVMIVELLNSAIEAVVDRIGSEYHELSGRAKDLGSAAVLIAIIDAVITWC
ILLWSHFG
1|30|47|2|52|70|3|99|117

3o7q_A|438|12|3.14|X-RAY|0.53|288.84|56.35
MGNTSIQTQSYRAVDKDAGQSRSYIIPFALLCSLFFLWAVANNLNDILLPQFQQAFTLTNFQAGLIQSAFYFGYFIIPIPAGILMKKLSYKAGIITGLFLYALGAALFWPAAEIMNYTLFLV
GLFIIAAGLGCLETAANPFVTVLGPESSGHFRLNLAQTFNSFGAIIAVVFGQSLILSNVPHQSQDVLDKMSPEQLSAYKHSLVLSVQTPYMIIVAIVLLVALLIMLTKFPALQSDNHSDAKQ
GSFSASLSRLARIRHWRWAVLAQFCYVGAQTACWSYLIRYAVEEIPGMTAGFAANYLTGTMVCFFIGRFTGTWLISRFAPHKVLAAYALIAMALCLISAFAGGHVGLIALTLCSAFMSIQYP
TIFSLGIKNLGQDTKYGSSFIVMTIIGGGIVTPVMGFVSDAAGNIPTAELIPALCFAVIFIFARFRSQTATN
1|28|49|2|64|85|3|90|110|4|118|142|5|152|177|6|210|228|7|260|281|8|300|320|9|324|344|10|348|373|11|381|405|12|412|430

4tsy_A|179|1|3.14|X-RAY|0.79|62.33|5.07
SADVAGAVIDGAGLGFDVLKTVLEALGNVKRKIAVGIDNESGKTWTAMNTYFRSGTSDIVLPHKVAHGKALLYNGQKNRGPVATGVVGVIAYSMSDGNTLAVLFSVPYDYNWYSNWWNVRVY
KGQKRADQRMYEELYYHRSPFRGDNGWHSRGLGYGLKSRGFMNSSGHAILEIHVTKA
1|6|26

3hfx_A|504|14|3.15|X-RAY|0.5|345.53|75.13
MKNEKRKTGIEPKVFFPPLIIVGILCWLTVRDLDAANVVINAVFSYVTNVWGWAFEWYMVVMLFGWFWLVFGPYAKKRLGNEPPEFSTASWIFMMFASCTSAAVLFWGSIEIYYYISTPPFG
LEPNSTGAKELGLAYSLFHWGPLPWATYSFLSVAFAYFFFVRKMEVIRPSSTLVPLVGEKHAKGLFGTIVDNFYLVALIFAMGTSLGLATPLVTECMQWLFGIPHTLQLDAIIITCWIILNA
ICVACGLQKGVRIASDVRSYLSFLMLGWVFIVSGASFIMNYFTDSVGMLLMYLPRMLFYTDPIAKGGFPQGWTVFYWAWWVIYAIQMSIFLARISRGRTVRELCFGMVLGLTASTWILWTVL
GSNTLLLIDKNIINIPNLIEQYGVARAIIETWAALPLSTATMWGFFILCFIATVTLVNACSYTLAMSTCREVRDGEEPPLLVRIGWSILVGIIGIVLLALGGLKPIQTAIIAGGCPLFFVNI
MVTLSFIKDAKQNWKD
1|13|30|2|52|67|3|92|98|4|102|112|5|132|142|6|144|151|7|187|219|8|231|248|9|254|277|10|312|335|11|350|372|12|404|431|13|44
7|466|14|469|490

1rh5_C|53|1|3.2|X-RAY|0.54|32.41|3.56
MSKREETGLATSAGLIRYMDETFSKIRVKPEHVIGVTVAFVIIEAILTYGRFL
1|32|50

5vb2_B|285|6|3.2|X-RAY|0.5|194.41|29.35
****
****
****
****
****
****
****
****
****
****
****
****
****
****
****
****
******************************MDYKDDDDKGSLVPRGSHMYLRITNIVESSFFTKFIIYLIVLNGITMGLETSKTFMQSFGVYTTLFNQIVITIFTIEIILRIYVHRISFFKD
PWSLFDFFVVAISLVPTSSGFEILRVLRVLRLFRLVTAVPQMRKIVSALISVIPGMLSVIALMTLFFYIFAIMATQLFGERFPEWFGTLGESFYTLFQVMTLESWSMGIVRPLMEVYPYAWV
FFIPFIFVVFFVMINLYVAIIVDAMAILNQKEEQHIIDEVQSHEDNINNEIIKLREEIVELKELIKTSLKN
1|2013|2033|2|2044|2070|3|2071|2091|4|2097|2118|5|2129|2151|6|2193|2217

3jcu_K|59|1|3.2|E-MIC|0.51|38.22|3.53
MLNIFSLICLNSALYSSSFFFGKLPEAYAFLSPIVDFMPVIPLFFFLLAFVWQAAVSFR
1|30|55

1q90_N|31|1|3.1|X-RAY|0.59|16.82|2.3
***GEPAIVQIGWAATCVMFSFSLSLVVWGRSGL
1|74|95

3wvf_A|551|5|3.2|X-RAY|0.73|240.57|30.71
MDSQRNLLVIALLFVSFMIWQAWEQDKNPQPQAQQTTQTTTTAAGSAADQGVPASGQGKLISVKTDVLDLTINTRGGDVEQALLPAYPKELNSTQPFQLLETSPQFIYQAQSGLTGRDGPDN
PANGPRPLYNVEKDAYVLAEGQNELQVPMTYTDAAGNTFTKTFVLKRGDYAVNVNYNVQNAGEKPLEISTFGQLKQSITLPPHLDTGSSNFALHTFRGAAYSTPDEKYEKYKFDTIADNENL
NISSKGGWVAMLQQYFATAWIPHNDGTNNFYTANLGNGIAAIGYKSQPVLVQPGQTGAMNSTLWVGPEIQDKMAAVAPHLDLTVDYGWLWFISQPLFKLLKWIHSFVGNWGFSIIIITFIVR
GIMYPLTKAQYTSMAKMRMLQPKIQAMRERLGDDKQRISQEMMALYKAEKVNPLGGCFPLLIQMPIFLALYYMLMGSVELRQAPFALWIHDLSAQDPYYILPILMGVTMFFIQKMSPTTVTD
PMQQKIMTFMPVIFTVFFLWFPSGLVLYYIVSNLVTIIQQQLIYRGLEKRGLLESSGENLYFQ
1|355|377|2|422|440|3|464|477|4|494|511|5|512|526

3wo7_A|254|5|3.2|X-RAY|0.52|168.52|25.18
**********GQDPITSESEGIWNHFFVYPMSWLITTVANLLNGSYGLSIIIVTILIRLALLPLTLKQQKSMRAMQVIRPEMEAIQKKYKEKASKDPKVQQEMQKELL
GLYQKHGVNPMAGCLPLFIQLPILMAFYFAIMRTEEIRYHTFLWFDLGQPDYILPFVAGITTYFQFKMTMSHQQQMQKTNPSDSDNPMANMMQMQMKVMLYVMPVMIIIAGLSLPSALSLYW
VIGNIFMIIQTYFIVVKAPPLEVLESSGENLYFQ
1|61|83|2|133|153|3|174|191|4|218|239|5|240|258

5lev_A|409|10|3.2|X-RAY|0.5|279.82|48.04
SMWAFSELPMPLLINLIVSLLGFVATVTLIPAFRGHFIAARLCGQDLNKTSRQQIPESQGVISGAVFLIILFCFIPFPFLNCFVKEQCKAFPHHEFVALIGALLAICCMIFLGFADDVLNLR
WRHKLLLPTAASLPLLMVYFTNFGNTTIVVPKPFRPILGLHLDLGILYYVYMGLLAVFCTNAINILAGINGLEAGQSLVISASIIVFNLVELEGDCRDDHVFSLYFMIPFFFTTLGLLYHNW
YPSRVFVGDTFCYFAGMTFAGVGILGHFSKTMLLFFMPQVFNFLYSLPQLLHIIPCPRHRIPRLNIKTGKLEMSYSKFKTKSLSFLGTFILKVAESLQLVTVHQSETEDGEFTECNNMTLIN
LLLKVLGPIHERNLTLLLLLLQILGSAITFSIRYQLVRLFYDV
1|10|29|2|58|79|3|94|116|4|122|143|5|166|187|6|193|213|7|219|241|8|252|271|9|272|291|10|379|399

4tqu_N|305|6|3.2|X-RAY|0.52|203.35|30.58
MLATPFYSRSDRIFGIVNAVLLGIFALCALYPIIYIFSMSISSGAAVTQGRVFLLPVDIDFSAYGRVLHDKLFWTSYANTIFYTVFGVVTSLIFIVPGAYALSKPRIRGRRVFGFIIAFTMW
FNAGMIPFFLNMRDLGLLDNRFGILIGFACNAFNIILMRNYFESISASFEEAARMDGANDLQILWKVYIPLAKPALATITLLCAISRWNGYFWAMVLLRAEEKIPLQVYLKKTIVDLNVNEE
FAGALLTNSYSMETVVGAIIVMSIIPVIIVYPVVQKYFTKGVMLGGVKELEHHHHHHHHHH
1|13|41|2|76|102|3|107|122|4|154|162|5|195|216|6|261|282

5f15_A|578|13|3.2|X-RAY|0.51|392.23|71.67
SYVMPQPTSQQRASVASSQSTQGAVGWSAATGWVVLFVAVALVVWFVSLDMRHLVGPDEGRYAEISREMFASGDWVTIRYNALKYFEKPPFHMWVTVVGYELFGLGEWQARLAVALSGLLGI
GVSMMAARRWFGARAAAFTGLALLAAPMWSVAAHFNTLDMTLAGVMSCVLAFMLMGQHPDASVAARRGWMVACWAAMGVAILTKGLVGIALPGLVLVVYTLVTRDWGLWRRLHLALGVVVML
VITVPWFYLVSVRNPEFPNFFFIHEHWQRYTSNIHSRSGSVFYFLPLVIGGFLPWAGIFPKLWTAMRAPVEGTQARFRPALMAGIWAIAIFVFFSISRSKLPGYIVPVIPALGILAGVALDR
LSPRSWGKQLIGMAIVAACGLLASPVVATLNANHIPNSFYRAYAVWVAVAFVVMLLGIAVARLLLRRGVLPSVAVYAMGMYLGFTVALLGHETVGRPASGADIAPQIAQKLTPEMPLYGVQM
LDHTLPFYLRHPLMMVGQADELTFGATVEPQRVVPDVDSFTKLWKNGQPAMAVMSPDTYLALAPTLSMYVVARDWRRVVVANVASLAGPQ
1|28|47|2|105|126|3|135|153|4|158|175|5|186|203|6|205|221|7|233|250|8|282|304|9|321|337|10|343|358|11|369|389|12|406|427|1
3|434|453

3mk7_C|311|2|3.2|X-RAY|0.76|122.23|11.19
MSTFWSGYIALLTLGTIVALFWLIFATRKGESAGTTDQTMGHAFDGIEEYDNPLPRWWFLLFIGTLVFGILYLVLYPGLGNWKGVLPGYEGGWTQEKQWEREVAQADEKYGPIFAKYAAMSV
EEVAQDPQAVKMGARLFANYCSICHGSDAKGSLGFPNLADQDWRWGGDAASIKTSILNGRIAAMPAWGQAIGEEGVKNVAAFVRKDLAGLPLPEGTDADLSAGKNVYAQTCAVCHGQGGEGM
AALGAPKLNSAAGWIYGSSLGQLQQTIRHGRNGQMPAQQQYLGDDKVHLLAAYVYSLSQKPEQLANQ
1|4|27|2|57|75

3jcu_E|83|1|3.2|E-MIC|0.63|43.55|4.16
MSGSTGERSFADIITSIRYWVIHSITIPSLFIAGWLFVSTGLAYDVFGSPRPNEYFTESRQGIPLITGRFDSLEQLDEFSRSF
1|19|38

2zjs_Y|434|10|3.2|X-RAY|0.51|295.13|54.34
MVKAFWSALQIPELRQRVLFTLLVLAAYRLGAFIPTPGVDLDKIQEFLRTAQGGVFGIINLFSGGNFERFSIFALGIMPYITAAIIMQILVTVVPALEKLSKEGEEGRRIINQYTRIGGIAL
GAFQGFFLATAFLGAEGGRFLLPGWSPGPFFWFVVVVTQVAGIALLLWMAERITEYGIGNGTSLIIFAGIVVEWLPQILRTIGLIRTGEVNLVAFLFFLAFIVLAFAGMAAVQQAERRIPVQ
YARKVVGGRVYGGQATYIPIKLNAAGVIPIIFAAAILQIPIFLAAPFQDNPVLQGIANFFNPTRPSGLFIEVLLVILFTYVYTAVQFDPKRIAESLREYGGFIPGIRPGEPTVKFLEHIVSR
LTLWGALFLGLVTLLPQIIQNLTGIHSIAFSGIGLLIVVGVALDTLRQVESQLMLRSYEGFLSRGRLR
1|18|33|2|77|90|3|113|135|4|151|175|5|183|204|6|214|235|7|272|290|8|309|326|9|369|388|10|398|412

3mk7_B|203|1|3.2|X-RAY|0.79|71.32|5.19
MKSHEKLEKNVGLLTLFMILAVSIGGLTQIVPLFFQDSVNEPVEGMKPYTALQLEGRDLYIREGCVGCHSQMIRPFRAETERYGHYSVAGESVYDHPFLWGSKRTGPDLARVGGRYSDDWHR
AHLYNPRNVVPESKMPSYPWLVENTLDGKDTAKKMSALRMLGVPYTEEDIAGARDSVNGKTEMDAMVAYLQVLGTALTNKR
1|12|35

3jcu_W|137|1|3.2|E-MIC|0.73|57.99|4.75
MATITASSSASLVARASLVHNSRVGVSSSPILGLPSMTKRSKVTCSIENKPSTTETTTTTNKSMGASLLAAAAAATISNPAMALVDERMSTEGTGLPFGLSNNLLGWILFGVFGLIWALYFV
YASGLEEDEESGLSL
1|103|124

4oh3_A|599|12|3.25|X-RAY|0.51|408.46|65.99
MSLPETKSDDILLDAWDFQGRPADRSKTGGWASAAMILCIEAVERLTTLGIGVNLVTYLTGTMHLGNATAANTVTNFLGTSFMLCLLGGFIADTFLGRYLTIAIFAAIQATGVSILTLSTII
PGLRPPRCNPTTSSHCEQASGIQLTVLYLALYLTALGTGGVKASVSGFGSDQFDETEPKERSKMTYFFNRFFFCINVGSLLAVTVLVYVQDDVGRKWGYGICAFAIVLALSVFLAGTNRYRF
KKLIGSPMTQVAAVIVAAWRNRKLELPADPSYLYDVDDIIAAEGSMKGKQKLPHTEQFRSLDKAAIRDQEAGVTSNVFNKWTLSTLTDVEEVKQIVRMLPIWATCILFWTVHAQLTTLSVAQ
SETLDRSIGSFEIPPASMAVFYVGGLLLTTAVYDRVAIRLCKKLFNYPHGLRPLQRIGLGLFFGSMAMAVAALVELKRLRTAHAHGPTVKTLPLGFYLLIPQYLIVGIGEALIYTGQLDFFL
RECPKGMKGMSTGLLLSTLALGFFFSSVLVTIVEKFTGKAHPWIADDLNKGRLYNFYWLVAVLVALNFLIFLVFSKWYVYKEKRLAEVGIELDDEPSIPMGHAAAGSLVPR
1|32|56|2|67|88|3|101|121|4|143|168|5|188|209|6|217|238|7|344|364|8|382|407|9|421|439|10|465|484|11|500|523|12|543|562

3ux4_A|201|6|3.26|X-RAY|0.54|127.7|24.48
MLGLVLLYVGIVLISNGICGLTKVDPKSTAVMNFFVGGLSIVCNVVVITYSALHPTAPVEGHHHHHHAEDIVQVSHHLTSFYGPATGLLFGFTYLYAAINHTFGLDWRPYSWYSLFVAINTV
PAAILSHYSDMLDDHKVLGITEGDWWAIIWLAWGVLWLTAFIENILKIPLGKFTPWLAIIEGILTAWIPAWLLFIQHWV
1|3|21|2|28|50|3|76|97|4|101|123|5|142|160|6|168|191

5nik_J|654|4|3.3|E-MIC|0.74|276.59|25.29
MTPLLELKDIRRSYPAGDEQVEVLKGISLDIYAGEMVAIVGASGSGKSTLMNILGCLDKATSGTYRVAGQDVATLDADALAQLRREHFGFIFQRYHLLSHLTAEQNVEVPAVYAGLERKQRL
LRAQELLQRLGLEDRTEYYPAQLSGGQQQRVSIARALMNGGQVILADEPTGALDSHSGEEVMAILHQLRDRGHTVIIVTHDPQVAAQAERVIEIRDGEIVRNPPAIEKVNVTGGTEPVVNTV
SGWRQFVSGFNEALTMAWRALAANKMRTLLTMLGIIIGIASVVSIVVVGDAAKQMVLADIRSIGTNTIDVYPGKDFGDDDPQYQQALKYDDLIAIQKQPWVASATPAVSQNLRLRYNNVDVA
ASANGVSGDYFNVYGMTFSEGNTFNQEQLNGRAQVVVLDSNTRRQLFPHKADVVGEVILVGNMPARVIGVAEEKQSMFGSSKVLRVWLPYSTMSGRVMGQSWLNSITVRVKEGFDSAEAEQQ
LTRLLSLRHGKKDFFTWNMDGVLKTVEKTTRTLQLFLTLVAVISLVVGGIGVMNIMLVSVTERTREIGIRMAVGARASDVLQQFLIEAVLVCLVGGALGITLSLLIAFTLQLFLPGWEIGFS
PLALLLAFLCSTVTGILFGWLPARNAARLDPVDALAREHHHHHH
1|270|294|2|521|544|3|573|601|4|611|632

2wcd_A|309|3|3.29|X-RAY|0.72|137.22|16.67
MHHHHHHTEIVADKTVEVVKNAIETADGALDLYNKYLDQVIPWQTFDETIKELSRFKQEYSQAASVLVGDIKTLLMDSQDKYFEATQTVYEWCGVATQLLAAYILLFDEYNEKKASAQKDIL
IKVLDDGITKLNEAQKSLLVSSQSFNNASGKLLALDSQLTNDFSEKSSYFQSQVDKIRKEAYAGAAAGVVAGPFGLIISYSIAAGVVEGKLIPELKNKLKSVQNFFTTLSNTVKQANKDIDA
AKLKLTTEIAAIGEIKTETETTRFYVDYDDLMLSLLKEAAKKMINTCNEYQKRHGKKTLFEVPEV
1|10|35|2|175|186|3|189|202

4rue_A|309|4|3.3|X-RAY|0.56|192.24|21.4
MTTAPQEPPARPLQAGSGAGPAPGRAMRSTTLLALLALVLLYLVSGALVFRALEQPHEQQAQRELGEVREKFLRAHPCVSDQELGLLIKEVADALGGGADPETQSTSQSSHSAWDLGSAFFF
SITIITTIGYGNVALRTDAGRLFCIFYALVGIPLFGILLAGVGDRLGSSLRHGIGHIEAIFLKWHVPPELVRVLSAMLFLLIGCLLFVLTPTFVFCYMEDWSKLEAIYFVIVTLTTVGFGDY
VAGADPRQDSPAYQPLVWFWILLGLAYFASVLTTIGNWLRVVSRRTRAEMGGLTAQSNSLEVLFQ
1|30|51|2|141|172|3|195|218|4|257|279

4gd3_S|335|1|3.3|X-RAY|0.91|58.99|5.77
LENKPRIPVVWIHGLECTCCTESFIRSAHPLAKDVILSLISLDYDDTLMAAAGTQAEEVFEDIITQYNGKYILAVEGNPPLGEQGMFCISSGRPFIEKLKRAAAGASAIIAWGTCASWGCVQ
AARPNPTQATPIDKVITDKPIIKVPGCPPIPDVMSAIITYMVTFDRLPDVDRMGRPLMFYGQRIHDKCYRRAHFDAGEFVQSWDDDAARKGYCLYKMGCKGPTTYNACSSTRWNDGVSFCIQ
SGHGCLGCAENGFWDRGSFYSRVVDIPQMGTHSTADTVGLTALGVVAAAVGVHAVASAVDQRRRHNQQPTETEHQPGNEDKQARSHHHHHH
1|279|293

4wis_A|735|10|3.3|X-RAY|0.58|444.27|61.87
MSNLKDFSQPGSGQESNFGVDFVIHYKVPAAERDEAEAGFVQLIRALTTVGLATEVRHGENESLLVFVKVASPDLFAKQVYRARLGDWLHGVRVSAPHNDIAQALQDEPVVEAERLRLIYLM
ITKPHNEGGAGVTPTNAKWKHVESIFPLHSHSFNKEWIKKWSSKYTLEQTDIDNIRDKFGESVAFYFAFLRSYFRFLVIPSAFGFGAWLLLGQFSYLYALLCGLWSVVFFEYWKKQEVDLAV
QWGVRGVSSIQQSRPEFEWEHEAEDPITGEPVKVYPPMKRVKTQLLQIPFALACVVALGALIVTCNSLEVFINEVYSGPGKQYLGFLPTIFLVIGTPTISGVLMGAAEKLNAMENYATVDAH
DAALIQKQFVLNFMTSYMALFFTAFVYIPFGHILHPFLNFWRATAQTLTFSEKELPTREFQINPARISNQMFYFTVTAQIVNFATEVVVPYIKQQAFQKAKQLKSGSKVQEDHEEEAEFLQR
VREECTLEEYDVSGDYREMVMQFGYVAMFSVAWPLAACCFLVNNWVELRSDALKIAISSRRPIPWRTDSIGPWLTALSFLSWLGSITSSAIVYLCSNSKNGTQGEASPLKAWGLLLSILFAE
HFYLVVQLAVRFVLSKLDSPGLQKERKERFQTKKRLLQENLGQDAAEEAAAPGIEHSEKITREALEEEARQASIRGHGTPEEMFWQRQRGMQETIEIGRRMIEQQLAAGKNGKKSAPAVPSE
KAS
1|195|211|2|218|233|3|285|314|4|325|354|5|371|395|6|430|456|7|505|519|8|523|535|9|561|584|10|600|621

5jrw_A|373|2|3.3|X-RAY|0.83|111.2|11.66
MGSSHHHHHHSSGRENLYFQGHVEEKRLSAKKGLPPGTLVYTGKYREDFEIEVMNYSIEEFREFKTTDVESVLPFRDSSTPTWINITGIHRTDVVQRVGEFFGIHPLVLERILNVHQRPKVE
FFENYVFIVLKMFTYDKNLHELESEQVSLILTKNCVLMFQEKIGDVFDPVRERIRYNRGIIRKKRADYLLYSLIDALVDDYFVLLEKIDDEIDVLEEEVLERPEKETVQRTHQLKRNLVELR
KTIWPLREVLSSLYRDVPPLIEKETVPYFRRVYDHTIQIADTVETFRDIVSGLLDVYLSSVSNKTNEVMKVLTIIATIFMPLTFIAGIYGMNFEYMPELRWKWGYPVVLAVMGVIAVIMVVY
FKKKKWL
1|293|305|2|325|345

5llu_A|522|12|3.32|X-RAY|0.51|354.29|67.08
MTKSNGEEPKMGGRMERFQQGVSKRTLLAKKKVQNITKEDVKSFLRRNALLLLTVLAVILGVVLGFLLRPYPLSPREVKYFAFPGELLMRMLKMLILPLIVSSLITGLASLDAKASGRLGMR
AVVYYMSTTIIAVVLGIILVLIIHPGAASAAITASVGAAGSAENAPSKEVLDCFLDLARNIFPSNLVSAAFRSYSTTYEERTITGTRVKVPVGQEVEGMNILGLVVFSIVFGIALGKMGEQG
QLLVDFFNSLNEATMKLVAIIMWYAPLGILFLIAGKIVEMEDLEVLGGQLGMYMVTVIVGLVIHGLIVLPLIYFLITRKNPFVFIAGILQALITALGTSSSSATLPITFKCLEENNGVDKRI
TRFVLPVGATINMDGTALYEAVAAIFIAQVNNYELDFGQIITISITATAASIGAAGIPQAGLVTMVIVLTAVGLPTDDITLIIAVDWLLDRFRTMVNVLGDALGAGIVEHLSRKELEKQDAE
LGNSVIEENEMKKPYQLIAQDNETEKPIDSETKM
1|49|67|2|79|105|3|119|145|4|223|236|5|254|281|6|298|320|7|325|342|8|343|357|9|366|376|10|380|389|11|406|418|12|452|474

4zwj_A|906|7|3.3|X-RAY|0.68|445.49|45.48
NIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGILRNAKLKPVYDSLDAVRRAALINMVFQMGETGVAGFTNSLRMLQQ
KRWDEAAVNLAKSRWYNQTPNRAKRVITTFRTGTWDAYMCGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVADL
FMVLGGFTSTLYTSLHGYFVFGPTGCNLQGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLAGWSRYIPEGLQCSCGIDYYTLKPEVNNESFVIY
MFVVHFTIPMIIIFFCYGQLVFTVKEAAAQQQESATTQKAEKEVTRMVIIYVIAFLICWVPYASVAFYIFTHQGSCFGPIFMTIPAFFAKSAAIYNPVIYIMMNKQFRNCMLTTICCGKNPL
GDDEASATVSKTETSQVAPAAAAGSAGSAGSAGSASHVIFKKVSRDKSVTIYLGKRDYVDHVSQVEPVDGVVLVDPELVKGKKVYVTLTCAFRYGQEDIDVMGLTFRRDLYFSRVQVYPPVG
AMSVLTQLQESLLKKLGDNTYPFLLTFPDYLPCSVMLQPAPQDVGKSCGVDFEVKAFASDITDPEEDKIPKKSSVRLLIRKVQHAPPEMGPQPSAEASWQFFMSDKPLNLSVSLSKEIYFHG
EPIPVTVTVTNNTDKVVKKIKVSVEQIANVVLYSSDYYVKPVASEETQEKVQPNSTLTKTLVLVPLLANNRERRGIALDGKIKHEDTNLASSTIIKEGIDRTVMGILVSYHIKVKLTVSGFL
GELTSSEVATEVPFRLMHPQPEDPAKESVQDENAAAEEFARQNLKDTGENTE
1|35|63|2|73|99|3|109|134|4|151|173|5|201|224|6|253|277|7|286|309

4hkr_B|214|4|3.35|X-RAY|0.52|141.68|18.76
****
**********MSQSGEDLHSPTYLSWRKLQLSRAKLKASSKTSALLSGFAMVAMVEVQLDHDTNVPPGMLIAFAICTTLLVAVHMLALMISTCILPNIETVSNLHSISLVHESPHERLHWYI
ETAWAFSTLLGLILFLLEIAILCWVKFYDLSRRAAWSATVVLIPVMIIFMAFAIHFYRSLVSHKYEVTVSGIRELEMLKEQMEQDHLEHHNNIRNNGEGEEF
1|160|180|2|193|212|3|249|271|4|276|298

5i31_A|906|12|3.35|X-RAY|0.59|536.79|74.65
****
****
**********RWGSFCVRNPGCVIFFSLVFITACSS
GLVFVRVTTNPVDLWSAPSSQARLEKEYFDQHFGPFFRTEQLIIRAPLTDKHIYQPYPSGADVPFGPPLDIQILHQVLDLQIAIENITASYDNETVTLQDICLAPLSPYNTNCTILSVLNYF
QNSHSVLDHKKGDDFFVYADYHTHFLYCVRAPASLNDTSLLHDPCLGTFGGPVFPWLVLGGYDDQNYNNATALVITFPVNNYYNDTEKLQRAQAWEKEFINFVKNYKNPNLTISFTAERSIE
DELNRESDSDVFTVVISYAIMFLYISLALGHMKSCRRLLVDSKVSLGIAGILIVLSSVACSLGVFSYIGLPLTLIVIEVIPFLVLAVGVDNIFILVQAYQRDERLQGETLDQQLGRVLGEVA
PSMFLSSFSETVAFFLGALSVMPAVHTFSLFAGLAVFIDFLLQITCFVSLLGLDIKRQEKNRLDIFCCVRGAEDGTSVQASESCLFRFFKNSYSPLLLKDWMRPIVIAIFVGVLSFSIAVLN
KVDIGLDQSLSMPDDSYMVDYFKSISQYLHAGPPVYFVLEEGHDYTSSKGQNMVCGGMGCNNDSLVQQIFNAAQLDNYTRIGFAPSSWIDDYFDWVKPQSSCCRVDNITDQFCNASVVDPAC
VRCRPLTPEGKQRPQGGDFMRFLPMFLSDNPNPKCGKGGHAAYSSAVNILLGHGTRVGATYFMTYHTVLQTSADFIDALKKARLIASNVTETMGINGSAYRVFPYSVFYVFYEQYLTIIDDT
IFNLGVSLGAIFLVTMVLLGCELWSAVIMCATIAMVLVNMFGVMWLWGISLNAVSLVNLVMSCGISVEFCSHITRAFTVSMKGSRVERAEEALAHMGSSVFSGITLTKFGGIVVLAFAKSQI
FQIFYFRMYLAMVLLGATHGLIFLPV
1|350|370|2|621|640|3|658|679|4|686|708|5|731|750|6|757|783|7|834|852|8|1096|1117|9|1123|1140|10|1147|1166|11|1188|1206|12
|1218|1243

5a63_A|709|1|3.4|E-MIC|0.94|95.6|6.53
MATAGGGSGADPGSRGLLRLLSFCVLLAGLCRGNSVERKIYIPLNKTAPCVRLLNATHQIGCQSSISGDTGVIHVVEKEEDLQWVLTDGPNPPYMVLLESKHFTRDLMEKLKGRTSRIAGLA
VSLTKPSPASGFSPSVQCPNDGFGVYSNSYGPEFAHCREIQWNSLGNGLAYEDFSFPIFLLEDENETKVIKQCYQDHNLSQNGSAPTFPLCAMQLFSHMHAVISTATCMRRSSIQSTFSINP
EIVCDPLSDYNVWSMLKPINTTGTLKPDDRVVVAATRLDSRSFFWNVAPGAESAVASFVTQLAAAEALQKAPDVTTLPRNVMFVFFQGETFDYIGSSRMVYDMEKGKFPVQLENVDSFVELG
QVALRTSLELWMHTDPVSQKNESVRNQVEDLLATLEKSGAGVPAVILRRPNQSQPLPPSSLQRFLRARNISGVVLADHSGAFHNKYYQSIYDTAENINVSYPEWLSPEEDLNFVTDTAKALA
DVATVLGRALYELAGGTNFSDTVQADPQTVTRLLYGFLIKANNSWFQSILRQDLRSYLGDGPLQHYIAVSSPTNTTYVVQYALANLTGTVVNLTREQCQDPSKVPSENKDLYEYSWVQGPLH
SNETDRLPRCVRSTARLARALSPAFELSQWSSTEYSTWTESRWKDIRARIFLIASKELELITLTVGFGILIFSLIVTYCINAKADVLFIAPREPGAVSY
1|669|690

4aw6_A|482|7|3.4|X-RAY|0.54|311.95|37.72
MGMWASLDALWEMPAEKRIFGAVLLFSWTVYLWETFLAQRQRRIYKTTTHVPPELGQIMDSETFEKSRLYQLDKSTFSFWSGLYSETEGTLILLFGGIPYLWRLSGRFCGYAGFGPEYEITQ
SLVFLLLATLFSALAGLPWSLYNTFVIEEKHGFNQQTLGFFMKDAIKKFVVTQCILLPVSSLLLYIIKIGGDYFFIYAWLFTLVVSLVLVTIYADYIAPLFDKFTPLPEGKLKEEIEVMAKS
IDFPLTKVYVVEGSKRSSHSNAYFYGFFKNKRIVLFDTLLEEYSVLNKDIQEDSGMEPRNEEEGNSEEIKAKVKNKKQGCKNEEVLAVLGHELGHWKLGHTVKNIIISQMNSFLCFFLFAVL
IGRKELFAAFGFYDSQPTLIGLLIIFQFIFSPYNEVLSFCLTVLSRRFEFQADAFAKKLGKAKDLYSALIKLNKDNLGFPVSDWLFSMWHYSHPPLLERLQALKTMKQHAENLYFQ
1|18|40|2|76|102|3|120|145|4|163|189|5|195|215|6|345|367|7|384|410

3hd7_B|109|1|3.4|X-RAY|0.69|51.08|4.49
****
**********GSHMDSSISKQALSEIETRHSEIIKLENSIRELHDMFMDMAMLVESQGEMIDRIEYNVEHAVDYV
ERAVSDTKKAVKYQSKARRKKIMIIICCVILGIIIASTIGGIFG
1|263|283

4q2e_A|290|9|3.4|X-RAY|0.51|195.19|43.11
MGSSHHHHHHSSGLVPRGSHMDDLKTRVITASVVAPFVVLCFVSYESLIGLVSAILILAGYELITLEMKERDARFFYVILLALYPVLYGLVFEEPTQPLSILFITGVVFSLITDKDPSQVFK
TVAAFSIALIYVTFFLSFFLPIYRDFGAANALLVLTSTWVFDSFAYFTGLKFGRTRISPRYSPRKSLEGVIGGFLGVVIYTFLYRLVVNDLLSVNVICFRTFLPFAATVAIMDTFGDIFECA
LKRHYGVKDSGKTLPGHGGMLDRIDGLLFVAPVSYIVFKILEGVVR
1|31|43|2|45|60|3|76|90|4|95|108|5|127|147|6|150|171|7|189|210|8|221|239|9|264|286

3rce_A|724|13|3.4|X-RAY|0.55|458.52|73.61
MELQQNFTDNNSIKYTCILILIAFAFSVLCRLYWVAWASEFYEFFFNDQLMITTNDGYAFAEGARDMIAGFHQPNDLSYFGSSLSTLTYWLYSILPFSFESIILYMSTFFASLIVVPIILIA
REYKLTTYGFIAALLGSIANSYYNRTMSGYYDTDMLVLVLPMLILLTFIRLTINKDIFTLLLSPIFIMIYLWWYPSSYSLNFAMIGLFGLYTLVFHRKEKIFYLAIALMIIALSMLAWQYKL
ALIVLLFAIFAFKEEKINFYMIWALIFISISILHLSGGLDPVLYQLKFYVFKASDVQNLKDAAFMYFNVNETIMEVNTIDPEVFMQRISSSVLVFILSFIGFILLLCKDHKSMLLALPMLALG
FMALRAGLRFTIYAVPVMALGFGYFLYAFFNFLEKKQIKLSLRNKNILLILIAFFSISPALMHIYYYKSSTVFTSYEASILNDLKNKAQREDYVVAWWDYGYPIRYYSDVKTLIDGGKHLGK
DNFFSSFVLSKEQIPAANMARLSVEYTEKSFKENYPDVLKAMVKDYNQTSAKDFLESLNDKNFKFDTNKTRDVYIYMPYRMLRIMPVVAQFANTNPDNGEQEKSLFFSQANAIAQDKTTGSV
MLDNGVEIINDFRALKVEGASIPLKAFVDIESITNGKFYYNEIDSKAQIYLLFLREYKSFVILDESLYNSAYIQMFLLNQYDQDLFEQVTNDTRAKIYRLKREFHHHHHHHHHH
1|13|33|2|105|125|3|128|148|4|156|174|5|179|196|6|199|217|7|222|235|8|240|256|9|263|280|10|336|352|11|355|370|12|376|396|1
3|412|432

5a63_B|467|8|3.4|E-MIC|0.56|290.68|44.77
MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRRSLGHPEPLSNGRPQGNSRQVVEQDEEEDEELTLKYGAKHVIMLFVPVTLCMVVVVATIKSVSFYTRKDGQLIYTPFTEDT
ETVGQRALHSILNAAIMISVIVVMTILLVVLYKYRCYKVIHAWLIISSLLLLFFFSFIYLGEVFKTYNVAVDYITVALLIWNFGVVGMISIHWKGPLRLQQAYLIMISALMALVFIKYLPEW
TAWLILAVISVYDLVAVLCPKGPLRMLVETAQERNETLFPALIYSSTMVWLVNMAEGDPEAQRRVSKNSKYNAESTERESQDTVAENDDGGFSEEWEAQRDSHLGPHRSTPESRAAVQELSS
SILAGEDPEERGVKLGLGDFIFYSVLVGKASATASGDWNTTIACFVAILIGLCLTLLLLAIFKKALPALPISITFGLVFYFATDYLVQPFMDQLAFHQFYI
1|85|102|2|171|186|3|195|214|4|221|240|5|243|260|6|383|399|7|403|421|8|436|458

5h36_A|215|7|3.41|X-RAY|0.6|123.54|26.51
MFQTVTILLDWFGLCIFTVTGALVASRKEMDIAGFVLLGAVTGVGGGTIRDLVLGRTPVFWVEEPAYVLACLGVAVFTFFFAHIPQSRYRFLLWLDAVGLSLFAVTGAERALQTGAGPVIAI
AMGVATATFGGILRDLLGGESPVILRREIYITAALLGAAAFVALDAFGAPRELALGAGFAAAFLSRAAGLVWGLSLPRYRARLESSGENLYFQ
1|5|27|2|32|54|3|64|85|4|88|112|5|118|140|6|152|169|7|173|194

3wak_A|875|13|3.41|X-RAY|0.57|538.78|80.2
MQNAESWFKKYWHLSVLVIAALISVKLRILNPWNSVFTWTVRLGGNDPWYYYRLIENTIHNFPHRIWFDPFTYYPYGSYTHFGPFLVYLGSIAGIIFSATSGESLRAVLAFIPAIGGVLAIL
PVYLLTREVFDKRAAVIAAFLIAIVPGQFLQRSILGFNDHHIWEAFWQVSALGTFLLAYNRWKGHDLSHNLTARQMAYPVIAGITIGLYVLSWGAGFIIAPIILAFMFFAFVLAGFVNADRK
NLSLVAVVTFAVSALIYLPFAFNYPGFSTIFYSPFQLLVLLGSAVIAAAFYQIEKWNDVGFFERVGLGRKGMPLAVIVLTALIMGLFFVISPDFARNLLSVVRVVQPKGGALTIAEVYPFFF
THNGEFTLTNAVLHFGALFFFGMAGILYSAYRFLKRRSFPEMALLIWAIAMFIALWGQNRFAYYFAAVSAVYSALALSVVFDKLHLYRALENAIGARNKLSYFRVAFALLIALAAIYPTYIL
ADAQSSYAGGPNKQWYDALTWMRENTPDGEKYDEYYLQLYPTPQSNKEPFSYPFETYGVISWWDYGHWIEAVAHRMPIANPFQAGIGNKYNNVPGASSFFTAENESYAEFVAEKLNVKYVVS
DIEMETGKYYAMAVWAEGDLPLAEKYYGGYFYYSPTGTFGYANSQWDIPLNSIIIPLRIPSELYYSTMEAKLHLFDGSGLSHYRMIYESDYPAEWKSYSSQVNLNNESQVLQTALYEAVMRA
RYGVSPTMGTQEVLYKYAYTQLYEKKMGIPVKIAPSGYVKIFERVKGAVVTGKVSANVTEVSVNATIKTNQNRTFEYWQTVEVKNGTYTVVLPYSHNSDYPVKPITPYHIKAGNVVKEITIY
ESQVQNGEIIQLDLELALVPR
1|11|31|2|111|135|3|136|146|4|163|182|5|197|217|6|218|235|7|245|266|8|278|297|9|314|336|10|375|400|11|405|423|12|427|447|1
3|469|493

5x5y_G|355|6|3.46|X-RAY|0.51|238.64|31.35
MVKLDRYIGVTVFVAILAVLGVILGLALLFAFIDELNDISASYGIGDALRFIFLTAPRRAYDMLPMAALIGCLVGLGTLASNSELTIMRAAGVSLSRIVWAVMKPMLVLMLAGILVGEYVAP
WTENIAQSGRALAQGGGDSQSSKRGLWHRQGREYIHINAVQPNGVLYGVTRYRFDEQRGLESASFAKRARFETDHWQLEEVTTTLLHPREKRSEVVKLPTERWDAQLSPQLLNTVVMEPEAL
SISGLWQYIHYLADQGLNNNRYWLAFWTKVLQPLVTAALVLMAISFIFGPLRSVTLGQRIFTGVLVGFVFRIAQDLLGPSSLVFDFPPLLAVVIPASICALAGVWLLRRAG
1|9|36|2|52|75|3|99|125|4|267|291|5|300|326|6|332|350

4yzf_A|911|14|3.5|X-RAY|0.57|558.5|87.25
MEELQDDYEDMMEENLEQEEYEDPDIPESQMEEPAAHDTEATATDYHTTSHPGTHKVYVELQELVMDEKNQELRWMEAARWVQLEENLGENGAWGRPHLSHLTFWSLLELRRVFTKGTVLLD
LQETSLAGVANQLLDRFIFEDQIRPQDREELLRALLLKHSHAGELEALGGVKPAVLTRSGDPSQPLLPQHSSLETQLFCEQGDGGTEGHSPSGILEKIPPDSEATLVLVGRADFLEQPVLGF
VRLQEAAELEAVELPVPIRFLFVLLGPEAPHIDYTQLGRAAATLMSERVFRIDAYMAQSRGELLHSLEGFLDCSLVLPPTDAPSEQALLSLVPVQRELLRRRYQSSPAKPDSSFYKGLDLNG
GPDDPLQQTGQLFGGLVRDIRRRYPYYLSDITDAFSPQVLAAVIFIYFAALSPAITFGGLLGEKTRNQMGVSELLISTAVQGILFALLGAQPLLVVGFSGPLLVFEEAFFSFCETNGLEYIV
GRVWIGFWLILLVVLVVAFEGSFLVRFISRYTQEIFSFLISLIFIYETFSKLIKIFQDHPLQKTYNYNVLMVPKPQGPLPNTALLSLVLMAGTFFFAMMLRKFKNSSYFPGKLRRVIGDFGV
PISILIMVLVDFFIQDTYTQKLSVPDGFKVSNSSARGWVIHPLGLRSEFPIWMMFASALPALLVFILIFLESQITTLIVSKPERKMVKGSGFHLDLLLVVGMGGVAALFGMPWLSATTVRSV
THANALTVMGKASTPGAAAQIQEVKEQRISGLLVAVLVGLSILMEPILSRIPLAVLFGIFLYMGVTSLSGIQLFDRILLLFKPPKYHPDVPYVKRVKTWRMHLFTGIQIICLAVLWVVKSTP
ASLALPFVLILTVPLRRVLLPLIFRNVELQCLDADDAKATFDEEEGRDEYDEVAMPV
1|403|427|2|437|459|3|467|475|4|486|510|5|519|541|6|572|593|7|601|623|8|661|691|9|702|719|10|731|738|11|760|777|12|785|805
|13|837|855|14|856|870

5tj6_A|1070|7|3.5|E-MIC|0.74|461.71|46.28
MASSSSTSCEPGDRQWYSFLASSLVTFGSGLVVIIIYRIVLWLCCRKKKCIQVSNPVPTARTTSLDQKSFMKNSDPEIGWMTEAKDWAGELISGQTTTGRILVGLVFLLSIASLIIYFIDAS
TNTSVETCLPWSSSTTQQVDLAFNVFFMIYFFIRFVAANDKLWFWVELFSFVDYFTIPPSFVAIYLDRNWLGLRFLRALRLMSIPDILTYLNVLKTSTLIRLVQLVVSFVSLWLTAAGFLHL
LENSGDPFFDFGNAQHLTYWECLYFLMVTMSTVGFGDIFATTVLGRTFVVIFIMIFIGLFASFIPEIAEILGKRQKYGGSYKKERGKRHVVVCGYITFDSVSNFLKDFLHKDREDVDVEIVF
LHKGLPGLELELGLLKRHFTQVEYFWGSVMDANDLERVKIQEADACLVLANKYCQDPDQEDAANIMRVISIKNYHSDIKVIVQLLQYHNKAYLLNIPSWDWKRGDDAVCVAELKLGFIAQSCL
APGFSTLMANLFTMRSYKPTPEMSQWQTDYMRGTGMEMYTEYLSSAFNALTFPEAAELCFSKLKLLLLAIEVRQEDTRESTLAINPGPKVKIENATQGFFIAESAEEVKRAFYYCKNCHANV
SDVRQIKKCKCRPLAMFKKGAAAVLALQRTPGLAVEPDGEANDKDKSRGTSTSKAVTSFPEKRKPQSRRKPSTTLKSKSPSEDSVPPPPPPVDEPRKFDSTGMFHWCPDRPLNDCLQDRSQA
SASGLRNHVVVCLFADAASPLIGLRNLVMPLRASNFHYHELKPTIIVGNLDYLHREWKTLQNFPKLSILPGSPLNRANLRAVNINLCDMCVIVSAKDRNMEDPNLVDKEAILCSLNIKAMTF
DDTMGLIQSSNFVPGGFSPLHENKRSQAGANVPLITELANDSNVQFLDQDDDDDPDTELYMTQPFACGTAFAVSVLDSLMSTSYFNDNALTLIRTLITGGATPELEQILAEGAGMRGGYCSP
AVLANRDRCRVAQISLFDGPLAQFGQGGHYGELFVYALRHFGILCIGLYRFRDTNESVRSPSSKRYVITNPPEDFPLLPTDQVYVLTYKQITNH
1|16|43|2|97|122|3|137|160|4|162|186|5|196|210|6|223|246|7|287|312

1fft_B|315|2|3.5|X-RAY|0.77|120.97|10.54
MRLRKYNKSLGWLSLFAGTVLLSGCNSALLDPKGQIGLEQRSLILTAFGLMLIVVIPAILMAVGFAWKYRASNKDAKYSPNWSHSNKVEAVVWTVPILIIIFLAVLTWKTTHALEPSKPLAH
DEKPITIEVVSMDWKWFFIYPEQGIATVNEIAFPANTPVYFKVTSNSVMNSFFIPRLGSQIYAMAGMQTRLHLIANEPGTYDGISASYSGPGFSGMKFKAIATPDRAAFDQWVAKAKQSPNT
MSDMAAFEKLAAPSEYNQVEYFSNVKPDLFADVINKFMAHGKSMDMTQPEGEHSAHEGMEGMDMSHAESAH
1|45|65|2|89|109

4p6v_B|415|10|3.5|X-RAY|0.51|280.22|53.12
MGLKKFLEDIEHHFEPGGKHEKWFALYEAAATLFYTPGLVTKRSSHVRDSVDLKRIMIMVWLAVFPAMFWGMYNAGGQAIAALNHLYSGDQLAAIVAGNWHYWLTEMLGGTMSSDAGWGSKM
LLGATYFLPIYATVFIVGGFWEVLFCMVRKHEVNEGFFVTSILFALIVPPTLPLWQAALGITFGVVVAKEVFGGTGRNFLNPALAGRAFLFFAYPAQISGDLVWTAADGYSGATALSQWAQG
GAGALINNATGQTITWMDAFIGNIPGSIGEVSTLALMIGAAFIVYMGIASWRIIGGVMIGMILLSTLFNVIGSDTNAMFNMPWHWHLVLGGFAFGMFFMATDPVSASFTNSGKWAYGILIGV
MCVLIRVVNPAYPEGMMLAILFANLFAPLFDHVVVERNIKRRLARYGKQ
1|56|76|2|120|147|3|158|170|4|176|194|5|204|217|6|272|290|7|295|312|8|335|345|9|356|374|10|382|397

4p6v_D|210|6|3.5|X-RAY|0.51|141.51|27.48
MSSAKELKKSVLAPVLDNNPIALQVLGVCSALAVTTKLETAFVMTLAVMFVTALSNFFVSLIRNHIPNSVRIIVQMAIIASLVIVVDQILKAYLYDISKQLSVFVGLIITNCIVMGRAEAFA
MKSEPIPSFIDGIGNGLGYGFVLMTVGFFRELLGSGKLFGLEVLPLISNGGWYQPNGLMLLAPSAFFLIGFMIWAIRTFKPEQVEAKE
1|28|36|2|39|62|3|72|90|4|100|119|5|128|156|6|184|198

2zw3_A|226|4|3.5|X-RAY|0.52|150.17|19.64
MDWGTLQTILGGVNKHSTSIGKIWLTVLFIFRIMILVVAAKEVWGDEQADFVCNTLQPGCKNVCYDHYFPISHIRLWALQLIFVSTPALLVAMHVAYRRHEKKRKFIKGEIKSEFKDIEEIK
TQKVRIEGSLWWTYTSSIFFRVIFEAAFMYVFYVMYDGFSMQRLVKCNAWPCPNTVDCFVSRPTEKTVFTVFMIAVSGICILLNVTELCYLLIRYCSGKSKKPV
1|22|45|2|73|93|3|136|157|4|187|212

4oaa_A|417|12|3.5|X-RAY|0.55|264.27|56.13
MYYLKNTNFWMFGLFFFFYFFIMGAYFPFFPIWLHDINHISKSDTWIIFAAISLFSLLFQPLFGLLSDKLGLRKYLLWIITGMLVMFAPFFIFIFGPLLQYNILVGSIVGGIYLGFCFNAGA
PAVEAFIEKVSRRSNFEFGRARMFGCVGWALCASIVGIMFTINNQFVFWLGSGCALILAVLLFFAKTDAPSSATVANAVGANHSAFSLKLALELFRQPKLWFLSLYVIGVSCTYDVFDQQFA
NFFTSFFATGEQGTRVFWYVTTMGELLNASIMFFAPLIINRIGGKNALLLAGTIMSVRIIGSSFATSALEVVILKTLHMFEVPFLLVGCFKYITSQFEVRFSATIYLVCFCFFKQLAMIFMS
VLAGNMYESIGFQGAYLVLGLVALGFTLISVFTLSGPGPLSLLRRQVNEVA
1|9|31|2|46|68|3|75|96|4|105|130|5|140|161|6|166|186|7|222|244|8|260|283|9|288|308|10|313|337|11|346|370|12|380|399

4p6v_F|408|1|3.5|X-RAY|0.9|83.2|5.96
MSTIIFGVVMFTLIILALVLVILFAKSKLVPTGDITISINGDPEKAIVTQPGGKLLTALAGAGVFVSSACGGGGSCGQCRVKIKSGGGDILPTELDHISKGEAREGERLACQVAVKADMDLE
LPEEIFGVKKWECTVISNDNKATFIKELKLAIPDGESVPFRAGGYIQIEAPAHHVKYADFDVPEKYRGDWDKFNLFRYESKVDEPIIRAYSMANYPEEFGIIMLNVRIATPPPNNPNVPPGQ
MSSYIWSLKAGDKCTISGPFGEFFAKDTDAEMVFIGGGAGMAPMRSHIFDQLKRLKSKRKMSYWYGARSKREMFYVEDFDGLAAENDNFVWHCALSDPQPEDNWTGYTGFIHNVLYENYLKD
HEAPEDCEYYMCGPPMMNAAVINMLKNLGVEEENILLDDFGG
1|5|26

5nmi_K|22|1|3.5|X-RAY|0.65|10.18|1.79
**************RNWVPTAQLWGAVGAVGLVSAT
1|18|34

5b2g_C|364|4|3.5|X-RAY|0.62|204.88|21.04
****
****
****
****
****
****
****
****
**********GSSGSSGNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGILRNAKLKPVYDSLDAVRRAA
LINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVNLAKSRWYNQTPNRAKRVITTFRTGTWDAYKGMASMGLQVMGIALAVLGWLAVMLCCALPMWRVTAFIGSNIVTSQTIWEGLWMNCVVQ
STGQMQCKVYDSLLALPQDLQAARALVIISIIVAALGVLLSVVGGKCTNCLEDESAKAKTMIVAGVVFLLAGLMVIVPVSWTAHNIIQDFYNPLVASGQKREMGASLYVGWAASGLLLLGGG
LLCCSGPSSGENLYFQ
1|6|26|2|76|101|3|116|141|4|162|182

3pjz_A|494|10|3.51|X-RAY|0.51|334.12|51.82
MQFRSIIRIVGLLLALFSVTMLAPALVALLYRDGAGVPFVTTFFVLLFCGAMCWFPNRRHKHELKSRDGFLIVVLFWTVLGSAGSLPFLIADNPNISVTDAFFESFSALTTTGATVIVGLDE
LPKAILFYRQFLQWFGGMGIIVLAVAILPVLGIGGMQLYRAEIPGPVKDTKMTPRIAETAKALWYIYLSLTIACAVAFWLAGMTPFDAISHSFSTIAIGGFSTHDASMGYFDSYAINLITVV
FLLISACNFTLHFAAFASGGVHPKYYWKDPEFRAFIFIQVLLFLVCFLLLLKHHSYTSPYDAFDQALFQTVSISTTAGFTTTGFADWPLFLPVLLLFSSFIGGCAGSTGGGMKVIRILLLTL
QGARELKRLVHPRAVYTIKVGGSALPQRVVDAVWGFFSAYALVFVVCMLGLIATGMDELSAFSAVAATLNNLGPGLGEVALHFGDVNDKAKWVLIVSMLFGRLEIFTLLILLTPTFWRSAAA
ENLYFQ
1|3|27|2|37|56|3|69|90|4|126|147|5|182|203|6|236|256|7|275|296|8|333|345|9|399|420|10|455|477

5vkq_A|1732|6|3.55|E-MIC|0.86|473.22|43.57
MSQPRGGRGGGRGGGVGRKTPSSLTGPPDESATPSERATPASKADSDPKDDSSSNGDKKDMDLFPAPKPPSAGASIRDTANKVLGLAMKSEWTPIEAELKKLEKYVANVGEDGNHIPLAGVH
DMNTGMTPLMYATKDNKTAIMDRMIELGADVGARNNDNYNVLHIAAMYSREDVVKLLLTKRGVDPFSTGGSRSQTAVHLVSSRQTGTATNILRALLAAAGKDIRLKADGRGKIPLLLAVESG
NQSMCRELLAAQTAEQLKATTANGDTALHLAARRRDVDMVRILVDYGTNVDTQNGEGQTPLHIAAAEGDEALLKYFYGVRASASIADNQDRTPMHLAAENGHAHVIEILADKFKASIFERTK
DGSTLMHIASLNGHAECATMLFKKGVYLHMPNKDGARSIHTAAAYGHTGIINTLLQKGEKVDVTTNDNYTALHIAVESAKPAVVETLLGFGADVHVRGGKLRETPLHIAARVKDGDRCALML
LKSGASPNLTTDDCLTPVHVAARHGNLATLMQLLEDEGDPLYKSNTGETPLHMACRACHPDIVRHLIETVKEKHGPDKATTYINSVNEDGATALHYTCQITKEEVKIPESDKQIVRMLLENG
ADVTLQTKTALETAFHYCAVAGNNDVLMEMISHMNPTDIQKAMNRQSSVGWTPLLIACHRGHMELVNNLLANHARVDVFDTEGRSALHLAAERGYLHVCDALLTNKAFINSKSRVGRTALHL
AAMNGFTHLVKFLIKDHNAVIDILTLRKQTPLHLAAASGQMEVCQLLLELGANIDATDDLGQKPIHVAAQNNYSEVAKLFLQQHPSLVNATSKDGNTCAHIAAMQGSVKVIEELMKFDRSGV
ISARNKLTDATPLQLAAEGGHADVVKALVRAGASCTEENKAGFTAVHLAAQNGHGQVLDVLKSTNSLRINSKKLGLTPLHVAAYYGQADTVRELLTSVPATVKSETPTGQSLFGDLGTESGM
TPLHLAAFSGNENVVRLLLNSAGVQVDAATIENGYNPLHLACFGGHMSVVGLLLSRSAELLQSQDRNGRTGLHIAAMHGHIQMVEILLGQGAEINATDRNGWTPLHCAAKAGHLEVVKLLCE
AGASPKSETNYGCAAIWFAASEGHNEVLRYLMNKEHDTYGLMEDKRFVYNLMVVSKNHNNKPIQEFVLVSPAPVDTAAKLSNIYIVLSTKEKERAKDLVAAGKQCEAMATELLALAAGSDSA
GKILQATDKRNVEFLDVLIENEQKEVIAHTVVQRYLQELWHGSLTWASWKILLLLVAFIVCPPVWIGFTFPMGHKFNKVPIIKFMSYLTSHIYLMIHLSIVGITPIYPVLRLSLVPYWYEVG
LLIWLSGLLLFELTNPSDKSGLGSIKVLVLLLGMAGVGVHVSAFLFVSKEYWPTLVYCRNQCFALAFLLACVQILDFLSFHHLFGPWAIIIGDLLKDLARFLAVLAIFVFGFSMHIVALNQS
FANFSPEDLRSFEKKNRNRGYFSDVRMHPINSFELLFFAVFGQTTTEQTQVDKIKNVATPTQPYWVEYLFKIVFGIYMLVSVVVLIQLLIAMMSDTYQRIQAQSDIEWKFGLSKLIRNMHRT
TTAPSPLNLVTTWFMWIVEKVKARMKKKKRPSLVQMMGIRQASPRTKAGAKWLSKIKKDSVALSQVHLSPLGSQASFSQANQNRIENVADWEAIAKKYRALVGDEEGGSLKDSDAESGSQEG
SGGQQPPAQVGRRAIKATLADTTK
1|1305|1321|2|1338|1356|3|1363|1387|4|1396|1420|5|1439|1462|6|1529|1554

5nmi_R|274|1|3.5|X-RAY|0.83|83.45|5.52
MLSVAARSGPFAPVLSATSRGVAGALRPLVQAAVPATSESPVLDLKRSVLCRESLRGQAAGRPLVASVSLNVPASVRYSHTDIKVPDFSDYRRPEVLDSTKSSKESSEARKGFSYLVTATTT
VGVAYAAKNVVSQFVSSMSASADVLAMSKIEIKLSDIPEGKNMAFKWRGKPLFVRHRTKKEIDQEAAVEVSQLRDPQHDLERVKKPEWVILIGVCTHLGCVPIANAGDFGGYYCPCHGSHYD
ASGRIRKGPAPLNLEVPSYEFTSDDMVIVG
1|34|59

5gjv_A|1873|24|3.6|E-MIC|0.59|inf|158.84
MEPSSPQDEGLRKKQPKKPLPEVLPRPPRALFCLTLQNPLRKACISIVEWKPFETIILLTIFANCVALAVYLPMPEDDNNSLNLGLEKLEYFFLTVFSIEAAMKIIAYGFLFHQDAYLRSGW
NVLDFIIVFLGVFTAILEQVNVIQSNTAPMSSKGAGLDVKALRAFRVLRPLRLVSGVPSLQVVLNSIFKAMLPLFHIALLVLFMVIIYAIIGLELFKGKMHKTCYYIGTDIVATVENEKPSP
CARTGSGRPCTINGSECRGGWPGPNHGITHFDNFGFSMLTVYQCITMEGWTDVLYWVNDAIGNEWPWIYFVTLILLGSFFILNLVLGVLSGEFTKEREKAKSRGTFQKLREKQQLEEDLRGY
MSWITQGEVMDVEDLREGKLSLEEGGSDTESLYEIEGLNKIIQFIRHWRQWNRVFRWKCHDLVKSRVFYWLVILIVALNTLSIASEHHNQPLWLTHLQDIANRVLLSLFTIEMLLKMYGLGL
RQYFMSIFNRFDCFVVCSGILELLLVESGAMTPLGISVLRCIRLLRLFKITKYWTSLSNLVASLLNSIRSIASLLLLLFLFIIIFALLGMQLFGGRYDFEDTEVRRSNFDNFPQALISVFQV
LTGEDWNSVMYNGIMAYGGPSYPGVLVCIYFIILFVCGNYILLNVFLAIAVDNLAEAESLTSAQKAKAEERKRRKMSRGLPDKTEEEKSVMAKKLEQKPKGEGIPTTAKLKVDEFESNVNEV
KDPYPSADFPGDDEEDEPEIPVSPRPRPLAELQLKEKAVPIPEASSFFIFSPTNKVRVLCHRIVNATWFTNFILLFILLSSAALAAEDPIRAESVRNQILGYFDIAFTSVFTVEIVLKMTTY
GAFLHKGSFCRNYFNILDLLVVAVSLISMGLESSTISVVKILRVLRVLRPLRAINRAKGLKHVVQCVFVAIRTIGNIVLVTTLLQFMFACIGVQLFKGKFFSCNDLSKMTEEECRGYYYVYK
DGDPTQMELRPRQWIHNDFHFDNVLSAMMSLFTVSTFEGWPQLLYRAIDSNEEDMGPVYNNRVEMAIFFIIYIILIAFFMMNIFVGFVIVTFQEQGETEYKNCELDKNQRQCVQYALKARPL
RCYIPKNPYQYQVWYVVTSSYFEYLMFALIMLNTICLGMQHYHQSEEMNHISDILNVAFTIIFTLEMILKLLAFKARGYFGDPWNVFDFLIVIGSIIDVILSEIDTFLASSGGLYCLGGGCG
NVDPDESARISSAFFRLFRVMRLIKLLSRAEGVRTLLWTFIKSFQALPYVALLIVMLFFIYAVIGMQMFGKIALVDGTQINRNNNFQTFPQAVLLLFRCATGEAWQEILLACSYGKLCDPES
DYAPGEEYTCGTNFAYYYFISFYMLCAFLIINLFVAVIMDNFDYLTRDWSILGPHHLDEFKAIWAEYDPEAKGRIKHLDVVTLLRRIQPPLGFGKFCPHRVACKRLVGMNMPLNSDGTVTFN
ATLFALVRTALKIKTEGNFEQANEELRAIIKKIWKRTSMKLLDQVIPPIGDDEVTVGKFYATFLIQEHFRKFMKRQEEYYGYRPKKDTVQIQAGLRTIEEEAAPEIRRTISGDLTAEEELER
AMVEAAMEERIFRRTGGLFGQVDTFLERTNSLPPVMANQRPLQFAEIEMEELESPVFLEDFPQDARTNPLARANTNNANANVAYGNSNHSNNQMFSSVHCEREFPGEAETPAAGRGALSHSH
RALGPHSKPCAGKLNGQLVQPGMPINQAPPAPCQQPSTDPPERGQRRTSLTGSLQDEAPQRRSSEGSTPRRPAPATALLIQEALVRGGLDTLAADAGFVTATSQALADACQMEPEEVEVAAT
ELLKARESVQGMASVPGSLSRRSSLGSLDQVQGSQETLIPPRP
1|52|70|2|85|107|3|115|137|4|162|183|5|195|219|6|307|332|7|433|453|8|459|484|9|495|516|10|523|542|11|558|581|12|632|659|13
|800|818|14|831|853|15|861|882|16|892|911|17|928|951|18|1039|1064|19|1119|1137|20|1146|1170|21|1181|1202|22|1231|1250|23|1
266|1289|24|1356|1380

3j2p_B|75|2|3.6|E-MIC|0.54|46.66|7.8
SVALVPHVGMGLETATETWMSSEGAWKHAQRIETWILRHPGFTIMAAILAYTIGTTHFQRALIFILLTAVAPSMT
1|39|52|2|58|70

3jc2_w|19|1|3.6|E-MIC|0.67|8.26|1.61
**********RLLLLLVVSNLLLCQGVVS
1|14|28

3j2p_A|495|2|3.6|E-MIC|0.88|118.61|12.28
MRCIGISNRDFVEGVSGGSWVDIVLEHGSCVTTMAKNKPTLDFELIETEAKQPATLRKYCIEAKLTNTTTDSRCPTQGEPSLNEEQDKRFVCKHSMVDRGWGNGCGLFGKGGIVTCAMFTCK
KNMKGKVVQPENLEYTIVITPHSGEEHAVGNDTGKHGKEIKITPQSSITEAELTGYGTVTMECSPRTGLDFNEMVLLQMENKAWLVHRQWFLDLPLPWLPGADTQGSNWIQKETLVTFKNPH
AKKQDVVVLGSQEGAMHTALTGATEIQMSSGNLLFTGHLKCRLRMDKLQLKGMSYSMCTGKFKVVKEIAETQHGTIVIRVQYEGDGSPCKIPFEIMDLEKRHVLGRLITVNPIVTEKDSPVN
IEAEPPFGDSYIIIGVEPGQLKLNWFKKGSSIGQMIETTMRGAKRMAILGDTAWDFGSLGGVFTSIGKALHQVFGAIYGAAFSGVSWIMKILIGVIITWIGMNSRSTSLSVSLVLVGVVTLY
LGVMVQA
1|454|468|2|475|492

5a40_B|128|4|3.6|X-RAY|0.7|59.36|13.11
MLTYAPLNFIAIGIGATLGAWLRWVLGLKLNGAGWPWGTLTANLVGGYLIGVMVALIASHPEWPAWIRLAAVTGFLGGLTTFSTFSAETVDMLCRGVYATAAAYAGASLAGSLAMTGLGLAT
VRLLLR
1|7|29|2|36|57|3|65|94|4|98|126

4wz7_L|89|3|3.6|X-RAY|0.53|56.25|10.67
**********MFIGTIILVLSFLGFVFNRRNIILAFICLETMLLGINLILLRNSVLFDDISGSLFAIVIIILAGVESAIGLSLLV
SYYRLRGVINSYGI**********
1|3|17|2|24|42|3|52|73

5vms_A|548|6|3.7|E-MIC|0.62|306.02|35.45
**********MATDPPRPTINLDPRVSIYSGRRPLLSRTNIQGRVYNFLERPTGWKCFVYHFTVFLI
VLICLIFSVLSTIQQYNNLATETLFWMEIVLVVFFGAEYVVRLWSAGCRSKYVGVWGRLRFARKPISVIDLIVVVASVIVLCVGSNGQVFATSAIRGIRFLQILRMLHVDRQGGTWRLLGSV
VFIHRQELITTLYIGFLGLIFSSYFVYLAEKDAIDSSGEYQFGSYADALWWGVVTVTTIGYGDKVPQTWIGKTIASCFSVFAISFFALPAGILGSGFALKVQQKQRQKHFNRQIPAAASLIQ
TAWRCYAAENPDSATWKIYIRKQSRNHHLMSPSPKPKKSAMVKKKKIRTERDEGSTDKMLNIPHITYDHVADDRKNDGYSVESYENTVRKPFGFLDPSTGPFIRTSSFTDDLDMEGDTLLTP
ITHISELKEHHRAAIKVIRRMQYFVAKKKFQQARKPYDVRDVIEQYSQGHLNLMVRIKELQRRLDQSLGKPSLFLSVSDKVKDKGINTIGSRLNRVEDKVTQMDHKLNLITDMLHHLLTNQQ
SNS
1|110|132|2|144|166|3|180|205|4|218|232|5|250|273|6|313|339

5va1_A|795|6|3.7|E-MIC|0.74|334.12|39.12
MPVRRGHVAPQNTFLDTIIRKFEGQSRKFIIANARVENCAVIYCNDGFCELCGYSRAEVMQRPCTCDFLHGPRTQRRAAAQIAQALLGAEERKVEIAFYRKDGSCFLCLVDVVPVKNEDGAV
IMFILNFEVVMEKDMVGSGADVLPEYKLQAPRIHRWTILHYSPFKAVWDWLILLLVIYTAVFTPYSAAFLLKETEEGPPATECGYACQPLAVVDLIVDIMFIVDILINFRTTYVNANEEVVS
HPGRIAVHYFKGWFLIDMVAAIPFDLLIFGSGSEELIGLLKTARLLRLVRVARKLDRYSEYGAAVLFLLMCTFALIAHWLACIWYAIGNMEQPHMDSRIGWLHNLGDQIGKPYNSSGLGGPS
IKDKYVTALYFTFSSLTSVGFGNVSPNTNSEKIFSICVMLIGSLMYASIFGNVSAIIQRLYSGTARYHTQMLRVREFIRFHQIPNPLRQRLEEYFQHAWSYTNGIDMNAVLKGFPECLQADI
CLHLNRSLLQHCKPFRGATKGCLRALAMKFKTTHAPPGDTLVHAGDLLTALYFISRGSIEILRGDVVVAILGKNDIFGEPLNLYARPGKSNGDVRALTYCDLHKIHRDDLLEVLDMYPEFSD
HFWSSLEITFNLRDTNMIPGGRQYQELPRCPAPTPSLLNIPLSSPGRRPRGDVESRLDALQRQLNRLETRLSADMATVLQLLQRQMTLVPPAYSAVTTPGPGPTSTSPLLPVSPLPTLTLDS
LSQVSQFMACEELPPGAPELPQEGPTRRLSLPGQLGALTSQPLHRHGSDPGSEASNSLEVLFQ**********
****
1|407|427|2|452|471|3|493|508|4|523|539|5|546|567|6|642|665

5eul_E|70|1|3.7|X-RAY|0.53|43.86|3.81
MQRVTNFFKEVVRELKKVSWPNRKELVNYTAVVLATVAFFTVFFAVIDLGISQLIRLVFEGGHHHHHHHH
1|30|55

5i6c_B|574|14|3.7|X-RAY|0.51|388.31|77.59
MDNSIHSTDGPDSVIPNSNPKKTVRQRVRLLARHLTTREGLIGDYDYGFLFRPELPFMKKDPRAPPFFGLNEKIPVLLAFILGLQHALAMLAGVVTPPLIISSSLSLPSDLQQYLVSTSLIV
CGLLSMVQITRFHIYKTPYYIGSGVLSVMGVSFSIISVASGAFNQMYSNGFCQLDEAGNRLPCPEAYGALIGTSACCALVEILLAFVPPKVIQKIFPPIVTGPTVMLIGISLIGTGFKDWAG
GSACMDDGMLCPSATAPRPLPWGSPEFIGLGFLVFVSIILCERFGAPIMKSCSVVIGLLVGCIVAAACGYFSHADIDAAPAASFIWVKTFPLSVYGPMVLPIIAVFIICACECIGDVTATCD
VSRLEVRGGTFESRIQGAVLADGINSVVAALATMTPMTTFAQNNVVIALTRCANRWAGYCCCLILIVAGIFAKFAAAIVAIPNSVMGGMKTFLFASVVISGQAIVAKAPFTRRNRFILTASM
ALGYGATLVPTWFGNVFPQTENRDLEGFENAIELVLETGFAVTAFVAMLLNAIMPAEVEEIGAVTPMPVSAHDNRDGEAEYQSKQA
1|76|101|2|113|131|3|155|164|4|189|207|5|221|239|6|272|288|7|295|309|8|342|361|9|381|398|10|407|413|11|422|437|12|450|473|
13|478|496|14|526|541

5eul_A|836|1|3.7|X-RAY|0.95|92.01|6.71
MLGILNKMFDPTKRTLNRYEKIANDIDAIRGDYENLSDDALKHKTIEFKERLEKGATTDDLLVEAFAVVREASRRVTGMFPFKVQLMGGVALHDGNIAEMKTGEGKTLTSTLPVYLNALTGK
GVHVVTVNEYLASRDAEQMGKIFEFLGLTVGLNLNSMSKDEKREAYAADITYSTNNELGFDYLRDNMVLYKEQMVQRPLHFAVIDEVDSILIDEARTPLIISGQAAKSTKLYVQANAFVRTL
KAEKDYTYDIKTKAVQLTEEGMTKAEKAFGIDNLFDVKHVALNHHINQALKAHVAMQKDVDYVVEDGQVVIVDSFTGRLMKGRRYSEGLHQAIEAKEGLEIQNESMTLATITFQNYFRMYEK
LAGMTGTAKTEEEEFRNIYNMQVVTIPTNRPVVRDDRPDLIYRTMEGKFKAVAEDVAQRYMTGQPVLVGTVAVETSELISKLLKNKGIPHQVLNAKNHEREAQIIEEAGQKGAVTIATNMAG
RGTDIKLGEGVKELGGLAVVGTERHESRRIDNQLRGRSGRQGDPGITQFYLSMEDELMRRFGAERTMAMLDRFGMDDSTPIQSKMVSRAVESSQKRVEGNNFDSRKQLLQYDDVLRQQREVI
YKQRFEVIDSENLREIVENMIKSSLERAIAAYTPREELPEEWKLDGLVDLINTTYLDEGALEKSDIFGKEPDEMLELIMDRIITKYNEKEEQFGKEQMREFEKVIVLRAVDSKWMDHIDAMD
QLRQGIHLRGSGGSGGKKTAIAIAVALAGFATVASYAQYEDGCSGELERQHTFAGGPGAQTNPLREYQMEGFAMFEHMIESIEDEVAKFVMKAEITSLEVLFQG
1|752|771

5tqq_A|671|14|3.76|E-MIC|0.52|446.72|81.17
**********MRVRRGIRGGLDWLKRKLFCVGEDWYFLTVLGVLMALISFTMSFTVGRVVRAHKWLYREIGDSHLLRYLSWTVYPVALVSFSSGFSQSITPFSGGSG
IPELKTILSGVVLEDYLDIKNFGAKAVGLTCTLASGSTIFLGKVGPFVHLSVMIAAYLGRVRAKATGESENKSKRNEMLVAGAAVGVATVFAAPFSGVLFCIEVVSSHFSVWDYWRGFFAAT
CGAFMFRLLAVFNSEQETITSLYKTSFRVEVPFDLPEIFFFVALGAICGVASCAYLFCQRKFLGFVKTNPVLSKLMATSKPLYSALAALVLASVTYPPGAGRFMASRLSMREYLDSLLDHNS
WALLTRQASPPWPVEPDPQNLWFEWYHPQFTIFGTLAFFLVMKFWMLILATTIPMPAGYFMPIFIFGAAIGRLLGEALSVAFPEGIVAGGVTNPIMPGGYALAGAAAFSGAVTHSISTALLA
FELTGQIVHALPVLMAVLAANAIAQSCQPSFYDGTIIVKKLPYLPWIRGRKISSHRVTVEHFMNRAITTLAKDTPQEEVVKVVTSTDMAEYPLVASTESQTLVGTMRRAQLVQALQAEPPSW
APGQQRCLQDILAEGCPVEPVTLKLSPETSLHQAHNLFELLNLQSLFVTSQGRAVGFVSWVELEKAISKLTNPPAPKSNSLEVLFQ
1|51|82|2|89|112|3|143|157|4|165|176|5|203|213|6|218|226|7|235|256|8|279|307|9|323|344|10|398|418|11|426|444|12|463|481|13
|482|492|14|497|515

5nj3_A|664|6|3.78|E-MIC|0.66|342.65|37.57
DYKDDDDKGSSSSNVEVFIPVSQGNTNGFPATASNDLKAFTEGAVLSFHNICYRVKLKSGFLPCRKPVEKEILSNINGIMKPGLNAILGPTGGGKSSLLDVLAARKDPSGLSGDVLINGAPR
PANFKCNSGYVVQDDVVMGTLTVRENLQFSAALRLATTMTNHEKNERINRVIQELGLDKVADSKVGTQFIRGVSGGERKRTSIGMELITDPSILFLDEPTTGLDSSTANAVLLLLKRMSKQG
RTIIFSIHQPRYSIFKLFDSLTLLASGRLMFHGPAQEALGYFESAGYHCEAYNNPADFFLDIINGDSTAVALNREEDFKATEIIEPSKQDKPLIEKLAEIYVNSSFYKETKAELHQLSGGEK
KKKITVFKEISYTTSFCHQLRWVSKRSFKNLLGNPQASIAQIIVTVVLGLVIGAIYFGLKNDSTGIQNRAGVLFFLTTNQCFSSVSAVELFVVEKKLFIHEYISGYYRVSSYFLGKLLSDLL
PMRMLPSIIFTCIVYFMLGLKPKADAFFVMMFTLMMVAYSASSMALAIAAGQSVVSVATLLMTICFVFMMIFSGLLVNLTTIASWLSWLQYFSIPRYGFTALQHNEFLGQNFCPGLNATGNN
PCNYATCTGEEYLVKQGIDLSPWGLWKNHVALACMIVIFLTIAYLKLLFLKKYS
1|392|413|2|428|448|3|470|497|4|505|528|5|534|556|6|626|650

3pjs_K|166|2|3.8|X-RAY|0.59|96.24|9.58
MHHHHHHPPMLSGLLARLVKLLLGRHGSALQWRAAGAATVLLVIVLLAGSYLAVLAERGAPGAQLITYPRALWWSVETATTVGYGDLYPVTLWGRLVAVVVMVAGITSFGLVTAALATWFVG
QEQQQQQQFVRHSEKAAEEAYTRTTRALHERFDRLERMLDDNRR
1|26|50|2|86|107

5sy1_A|670|9|3.9|E-MIC|0.58|405.04|51.89
MSAETVNNYDYSDWYENAAPTKAPVEVIPPCDPTADEGLFHICIAAISLVVMLVLAILARRQKLSDNQRGLTGLLSPVNFLDHTQHKGLAVAVYGVLFCKLVGMVLSHHPLPFTKEVANKEF
WMILALLYYPTLYYPLLACGTLHNKVGYVLGSLLSWTHFGILVWQKVDCPKTPQIYKYYALFGSLPQIACLAFLSFQYPLLLFKGLQNTETANASEDLSSSYYRDYVKKILKKKKPTKISSS
TSKPKLFDRLRDAVKSYIYTPEDVFRFPLKLAISVVVAFIALYQMALLLISGVLPTLHIVRRGVDENIAFLLAGFNIILSNDRQEVVRIVVYYLWCVEICYVSAVTLSCLVNLLMLMRSMVL
HRSNLKGLYRGDSLNVFNCHRSIRPSRPALVCWMGFTSYQAAFLCLGMAIQTLVFFICILFAVFLIIIPILWGTNLMLFHIIGNLWPFWLTLVLAALIQHVASRFLFIRKDGGTRDLNNRGS
LFLLSYILFLVNVMIGVVLGIWRVVITALFNIVHLGRLDISLLNRNVEAFDPGYRCYSHYLKIEVSQSHPVMKAFCGLLLQSSGQDGLSAQRIRDAEEGIQLVQQEKKQNKVSNAKRARAHW
QLLYTLVNNPSLVGSRKHFQCQSSESFINGALSRTSKEGSKKDGSVKEPNKEAESAAASN
1|36|57|2|88|107|3|120|142|4|148|170|5|181|202|6|273|295|7|343|364|8|402|429|9|452|468

2qfi_A|300|6|3.8|X-RAY|0.6|173.0|31.71
MNQSYGRLVSRAAIAATAMASLLLLIKIFAWWYTGSVSILAALVDSLVDIGASLTNLLVVRYSLQPADDNHSFGHGKAESLAALAQSMFISGSALFLFLTGIQHLISPTPMTDPGVGVIVTI
VALICTIILVSFQRWVVRRTQSQAVRADMLHYQSDVMMNGAILLALGLSWYGWHRADALFALGIGIYILYSALRMGYEAVQSLLDRALPDEERQEIIDIVTSWPGVSGAHDLRTRQSGPTRF
IQIHLEMEDSLPLVQAHMVADQVEQAILRRFPGSDVIIHQDPCSVVPREGKRSMLS
1|18|28|2|46|58|3|82|97|4|126|136|5|149|158|6|178|198

5lnk_w|125|1|3.9|E-MIC|0.7|57.29|4.63
ESSSSRAVIAPSTLAGKRPSEPTLRWQEDPEPEDENLYEKNPDSHGYDKDPAVDVWNMRVVFFFGFSIVLVLGSTFVAYLPDYRMQEWARREAERLVKYREAHGLPLMESNCFDPSKIQLPE
DED
1|56|78

5lnk_n|97|1|3.9|E-MIC|0.72|41.21|4.41
AHGHGHEHGPSKMELPDYKQWKIEGTPLETVQEKLAARGLRDPWGRNEAWRYMGGFANNVSFVGALLKGFKWGFAAFVVAVGAEYYLESQKKDKKHH
1|69|84

5gky_A|5037|6|3.8|E-MIC|0.95|567.41|50.31
MGDGGEGEDEVQFLRTDDEVVLQCSATVLKEQLKLCLAAEGFGNRLCFLEPTSNAQNVPPDLAICCFTLEQSLSVRALQEMLANTVEAGVESSQGGGHRTLLYGHAILLRHAHSRMYLSCLT
TSRSMTDKLAFDVGLQEDATGEACWWTMHPASKQRSEGEKVRVGDDLILVSVSSERYLHLSTASGELQVDASFMQTLWNMNPICSCCEEGYVTGGHVLRLFHGHMDECLTISAADSDDQRRL
VYYEGGAVCTHARSLWRLEPLRISWSGSHLRWGQPLRIRHVTTGRYLALTEDQGLVVVDACKAHTKATSFCFRVSKEKLDTAPKRDVEGMGPPEIKYGESLCFVQHVASGLWLTYAAPDPKA
LRLGVLKKKAILHQEGHMDDALFLTRCQQEESQAARMIHSTAGLYNQFIKGLDSFSGKPRGSGPPAGPALPIEAVILSLQDLIGYFEPPSEELQHEEKQSKLRSLRNRQSLFQEEGMLSLVL
NCIDRLNVYTTAAHFAEYAGEEAAESWKEIVNLLYELLASLIRGNRANCALFSTNLDWVVSKLDRLEASSGILEVLYCVLIESPEVLNIIQENHIKSIISLLDKHGRNHKVLDVLCSLCVCN
GVAVRSNQDLITENLLPGRELLLQTNLINYVTSIRPNIFVGRAEGSTQYGKWYFEVMVDEVVPFLTAQATHLRVGWALTEGYSPYPGGGEGWGGNGVGDDLYSYGFDGLHLWTGHVARPVTS
PGQHLLAPEDVVSCCLDLSVPSISFRINGCPVQGVFEAFNLDGLFFPVVSFSAGVKVRFLLGGRHGEFKFLPPPGYAPCHEAVLPRERLRLEPIKEYRREGPRGPHLVGPSRCLSHTDFVPC
PVDTVQIVLPPHLERIREKLAENIHELWALTRIEQGWTYGPVRDDNKRLHPCLVNFHSLPEPERNYNLQMSGETLKTLLALGCHVGMADEKAEDNLKKTKLPKTYMMSNGYKPAPLDLSHVR
LTPAQTTLVDRLAENGHNVWARDRVAQGWSYSAVQDIPARRNPRLVPYRLLDEATKRSNRDSLCQAVRTLLGYGYNIEPPDQEPSQVENQSRWDRVRIFRAEKSYTVQSGRWYFEFEAVTTG
EMRVGWARPELRPDVELGADELAYVFNGHRGQRWHLGSEPFGRPWQSGDVVGCMIDLTENTIIFTLNGEVLMSDSGSETAFREIEIGDGFLPVCSLGPGQVGHLNLGQDVSSLRFFAICGLQ
EGFEPFAINMQRPVTTWFSKSLPQFEPVPPEHPHYEVARMDGTVDTPPCLRLAHRTWGSQNSLVEMLFLRLSLPVQFHQHFRCTAGATPLAPPGLQPPAEDEARAAEPDPDYENLRRSAGGW
GEAEGGKEGTAKEGTPGGTPQPGVEAQPVRAENEKDATTEKNKKRGFLFKAKKAAMMTQPPATPALPRLPHDVVPADNRDDPEIILNTTTYYYSVRVFAGQEPSCVWVGWVTPDYHQHDMNF
DLSKVRAVTVTMGDEQGNVHSSLKCSNCYMVWGGDFVSPGQQGRISHTDLVIGCLVDLATGLMTFTANGKESNTFFQVEPNTKLFPAVFVLPTHQNVIQFELGKQKNIMPLSAAMFLSERKN
PAPQCPPRLEVQMLMPVSWSRMPNHFLQVETRRAGERLGWAVQCQDPLTMMALHIPEENRCMDILELSERLDLQRFHSHTLRLYRAVCALGNNRVAHALCSHVDQAQLLHALEDAHLPGPLR
AGYYDLLISIHLESACRSRRSMLSEYIVPLTPETRAITLFPPGRKGGNARRHGLPGVGVTTSLRPPHHFSPPCFVAALPAAGVAEAPARLSPAIPLEALRDKALRMLGEAVRDGGQHARDPV
GGSVEFQFVPVLKLVSTLLVMGIFGDEDVKQILKMIEPEVFTEEEEEEEEEEEEEEEEEDEEKEEDEEEEEKEDAEKEEEEAPEGEKEDLEEGLLQMKLPESVKLQMCNLLEYFCDQELQ
HRVESLAAFAERYVDKLQANQRSRYALLMRAFTMSAAETARRTREFRSPPQEQINMLLHFKDEADEEDCPLPEDIRQDLQDFHQDLLAHCGIQLEGEEEEPEEETSLSSRLRSLLETVRLVK
KKEEKPEEELPAEEKKPQSLQELVSHMVVRWAQEDYVQSPELVRAMFSLLHRQYDGLGELLRALPRAYTISPSSVEDTMSLLECLGQIRSLLIVQMGPQEENLMIQSIGNIMNNKVFYQHPN
LMRALGMHETVMEVMVNVLGGGETKEIRFPKMVTSCCRFLCYFCRISRQNQRSMFDHLSYLLENSGIGLGMQGSTPLDVAAASVIDNNELALALQEQDLEKVVSYLAGCGLQSCPMLLAKGY
PDIGWNPCGGERYLDFLRFAVFVNGESVEENANVVVRLLIRKPECFGPALRGEGGSGLLAAIEEAIRISEDPARDGPGVRRDRRREHFGEEPPEENRVHLGHAIMSFYAALIDLLGRCAPEM
HLIQAGKGEALRIRAILRSLVPLDDLVGIISLPLQIPTLGKDGALVQPKMSASFVPDHKASMVLFLDRVYGIENQDFLLHVLDVGFLPDMRAAASLDTATFSTTEMALALNRYLCLAVLPLI
TKCAPLFAGTEHRAIMVDSMLHTVYRLSRGRSLTKAQRDVIEDCLMALCRYIRPSMLQHLLRRLVFDVPILNEFAKMPLKLLTNHYERCWKYYCLPTGWANFGVTSEEELHLTRKLFWGIFD
SLAHKKYDQELYRMAMPCLCAIAGALPPDYVDASYSSKAEKKATVDAEGNFDPRPVETLNVIIPEKLDSFINKFAEYTHEKWAFDKIQNNWSYGENVDEELKTHPMLRPYKTFSEKDKEIYR
WPIKESLKAMIAWEWTIEKAREGEEERTEKKKTRKISQTAQTYDPREGYNPQPPDLSGVTLSRELQAMAEQLAENYHNTWGRKKKQELEAKGGGTHPLLVPYDTLTAKEKARDREKAQELLK
FLQMNGYAVTRGLKDMELDTSSIEKRFAFGFLQQLLRWMDISQEFIAHLEAVVSSGRVEKSPHEQEIKFFAKILLPLINQYFTNHCLYFLSTPAKVLGSGGHASNKEKEMITSLFCKLAALV
RHRVSLFGTDAPAVVNCLHILARSLDARTVMKSGPEIVKAGLRSFFESASEDIEKMVENLRLGKVSQARTQVKGVGQNLTYTTVALLPVLTTLFQHIAQHQFGDDVILDDVQVSCYRTLCSI

YSLGTTKNTYVEKLRPALGECLARLAAAMPVAFLEPQLNEYNACSVYTTKSPRERAILGLPNSVEEMCPDIPVLDRLMADIGGLAESGARYTEMPHVIEITLPMLCSYLPRWWERGPEAPPP
ALPAGAPPPCTAVTSDHLNSLLGNILRIIVNNLGIDEATWMKRLAVFAQPIVSRARPELLHSHFIPTIGRLRKRAGKVVAEEEQLRLEAKAEAEEGELLVRDEFSVLCRDLYALYPLLIRYV
DNNRAHWLTEPNANAEELFRMVGEIFIYWSKSHNFKREEQNFVVQNEINNMSFLTADSKSKMAKAGDAQSGGSDQERTKKKRRGDRYSVQTSLIVATLKKMLPIGLNMCAPTDQDLIMLAKT
RYALKDTDEEVREFLQNNLHLQGKVEGSPSLRWQMALYRGLPGREEDADDPEKIVRRVQEVSAVLYHLEQTEHPYKSKKAVWHKLLSKQRRRAVVACFRMTPLYNLPTHRACNMFLESYKAA
WILTEDHSFEDRMIDDLSKAGEQEEEEEEVEEKKPDPLHQLVLHFSRTALTEKSKLDEDYLYMAYADIMAKSCHLEEGGENGEAEEEEVEVSFEEKEMEKQRLLYQQSRLHTRGAAEMVLQM
ISACKGETGAMVSSTLKLGISILNGGNAEVQQKMLDYLKDKKEVGFFQSIQALMQTCSVLDLNAFERQNKAEGLGMVNEDGTVINRQNGEKVMADDEFTQDLFRFLQLLCEGHNNDFQNYLR
TQTGNTTTINIIICTVDYLLRLQESISDFYWYYSGKDVIEEQGKRNFSKAMSVAKQVFNSLTEYIQGPCTGNQQSLAHSRLWDAVVGFLHVFAHMMMKLAQDSSQIELLKELLDLQKDMVVM
LLSLLEGNVVNGMIARQMVDMLVESSSNVEMILKFFDMFLKLKDIVGSEAFQDYVTDPRGLISKKDFQKAMDSQKQFTGPEIQFLLSCSEADENEMINFEEFANRFQEPARDIGFNVAVLLT
NLSEHVPHDPRLRNFLELAESILEYFRPYLGRIEIMGASRRIERIYFEISETNRAQWEMPQVKESKRQFIFDVVNEGGEAEKMELFVSFCEDTIFEMQIAAQISEPEGEPEADEDEGMGEAA
AEGAEEGAAGAEGAAGTVAAGATARLAAAAARALRGLSYRSLRRRVRRLRRLTAREAATALAALLWAVVARAGAAGAGAAAGALRLLWGSLFGGGLVEGAKKVTVTELLAGMPDPTSDEVHG
EQPAGPGGDADGAGEGEGEGDAAEGDGDEEVAGHEAGPGGAEGVVAVADGGPFRPEGAGGLGDMGDTTPAEPPTPEGSPILKRKLGVDGEEEELVPEPEPEPEPEPEKADEENGEKEEVPEA
PPEPPKKAPPSPPAKKEEAGGAGMEFWGELEVQRVKFLNYLSRNFYTLRFLALFLAFAINFILLFYKVSDSPPGEDDMEGSAAGDLAGAGSGGGSGWGSGAGEEAEGDEDENMVYYFLEEST
GYMEPALWCLSLLHTLVAFLCIIGYNCLKVPLVIFKREKELARKLEFDGLYITEQPGDDDVKGQWDRLVLNTPSFPSNYWDKFVKRKVLDKHGDIFGRERIAELLGMDLASLEITAHNERKP
DPPPGLLTWLMSIDVKYQIWKFGVIFTDNSFLYLGWYMVMSLLGHYNNFFFAAHLLDIAMGVKTLRTILSSVTHNGKQLVMTVGLLAVVVYLYTVVAFNFFRKFYNKSEDEDEPDMKCDDMM
TCYLFHMYVGVRAGGGIGDEIEDPAGDEYELYRVVFDITFFFFVIVILLAIIQGLIIDAFGELRDQQEQVKEDMETKCFICGIGSDYFDTTPHGFETHTLEEHNLANYMFFLMYLINKDETE
HTGQESYVWKMYQERCWDFFPAGDCFRKQYEDQLS
1|4560|4579|2|4641|4662|3|4789|4807|4|4808|4822|5|4835|4853|6|4916|4941

5lnk_p|128|1|3.9|E-MIC|0.74|53.14|4.69
SFPKYEPSRLASLPTTLDPAEYDISSETRKAQAERLAIRSRLKREYQLQYNDPSRRGVVEDPALIRWTCARSANVYPNFRPNTKTSLLGALFGIGPLIFWYYVFKTDRDRKEKLIQEGKLDR
TFNISY
1|85|104

5lnk_v|158|1|3.9|E-MIC|0.78|57.66|4.93
ASHITKDMLPGPYPKTPEERAAAAKKYNMRVEDYEPYPDDGMGYGDYPKLPDRSQQERDPWYDWDHPDLRLNWGEPMHWDLDMYIRNRVDTSPTPVNWNLMCKHLFGFVAFMLFMFWVGETY
PTYQPVGPKQYPYNNLYLERGGDPNKEPEPVVHYEI
1|98|117

5lnk_u|72|1|3.9|E-MIC|0.6|40.28|3.97
AGGGAHIEPRYRQFPQLTRSQVIQAEFFSATMWFWILWRFWHDSDAVLGHFPYPDPSQWTDEELGIPPDDED
1|27|46

5lnk_N|347|11|3.9|E-MIC|0.55|220.65|49.32
MNPIILIIILMTVMLGTIIVMISTHWLLIWIGFEMNMLAIIPIMMKKHNPRATEASTKYFLTQSTASMLLMMAIIINLMFSGQWTVMKLFNPMASMLMTMALAMKLGMAPFHFWVPEVTQGI
PLSSGLILLTWQKLAPMSVLYQILPSINLDLILTLSILSITIGGWGGLNQTQLRKIMAYSSIAHMGWMTAVLLYNPTMTLLNLIIYIIMTSTMFTLFMANSTTTTLSLSHTWNKAPIMTILV
LITLLSMGGLPPLSGFMPKWMIIQEMTKNDSIILPTLMAITALLNLYFYMRLTYSTALTMFPSTNNMKMKWQFPTTKRMTLLPTMTVLSTMLLPLTPILSILE
1|3|22|2|27|45|3|54|75|4|96|123|5|124|142|6|152|171|7|175|194|8|200|220|9|239|256|10|277|298|11|326|343

5lnk_W|143|1|3.9|E-MIC|0.73|60.67|4.8
SGDHGKRLFIIKPSGFYDKRFLKLLRFYILLTGIPVVIGITLINVFIGEAELAEIPEGYVPEHWEYFKHPISRWIARTFFDAPEKNYERTMAILQIESEKAELRLKELEVRRLMRAKGDGPW
FQYPTIDKALIDHSPKATPDN
1|21|43

5lnk_m|83|1|3.9|E-MIC|0.62|44.65|4.14
APRVAAFLKNVWAKEPVLVASFAIAGLAVILPTLSPYTKYSLMINRATPYNYPVPLRDDGNMPDVPSHPQDPQGPSLEWLKRL
1|16|36

5lnk_q|143|1|3.9|E-MIC|0.76|55.53|4.82
AASKVKQDMPPVGGYGPIDYKRNLPRRGLSGYSMFAVGIGALLFGYWSMMRWNRERRRLQIEDFEARIALMPLLQAEKDRRVLQMLRENLEEEATIMKDVPGWKVGESVFHTTRWVTPMMGE
LYGLRTGEEILSSTYGFIWYT
1|31|50

5lnk_o|120|2|3.9|E-MIC|0.52|77.74|8.62
MMTGRQARAPLQFLPDEARSLPPPKLTDPRLAYIGFLGYCSGLIDNAIRRRPVLSAGLHRQFLYITSFVFVGYYLLKRQDYMYAVRDHDMFSYIKSHPEDFPEKDKKTYREVFEEFHPVR
1|26|48|2|53|76

5kyh_R|303|1|4.0|E-MIC|0.93|45.17|5.68
VSPVIATLLLILIAVAAAVLLYTWVSGLSANVAGTQVTGKSLTLIQATWARPATNVGTTISKDSFDRSKAVLILSFQPPAQVLQGGQAITIDAIDVLYQGRVVCHYDSFPMTADDKYHIGQT
IGGLTAFGLVFWGFVGSTLSDFDAHNETLVPGGIIHGKSDLATQPAARGYLGGDKGVTGEVHPGEKYKPDILLSFDDQYPFATILAGTWEVNYVSTNYVETNFRNTSAVIKFDRFVNTHYSP
DTQNNNGVPIFDVASASQSNFAVVIWCPNVNPNVMQSVDVKMVFSDGSTWEASVPLSIT
1|5|15

5xjy_A|2305|12|4.1|E-MIC|0.8|inf|85.83
MADYKDDDDKSGPDEVDASGRMACWPQLRLLLWKNLTFRRRQTCQLLLEVAWPLFIFLILISVRLSYPPYEQHECHFPNKAMPSAGTLPWVQGIICNANNPCFRYPTPGEAPGVVGNFNKSI
VARLFSDARRLLLYSQKDTSMKDMRKVLRTLQQIKKSSSNLKLQDFLVDNETFSGFLYHNLSLPKSTVDKMLRADVILHKVFLQGYQLHLTSLCNGSKSEEMIQLGDQEVSELCGLPREKLA
AAERVLRSNMDILKPILRTLNSTSPFPSKELAEATKTLLHSLGTLAQELFSMRSWSDMRQEVMFLTNVNSSSSSTQIYQAVSRIVCGHPEGGGLKIKSLNWYEDNNYKALFGGNGTEEDAET
FYDNSTTPYCNDLMKNLESSPLSRIIWKALKPLLVGKILYTPDTPATRQVMAEVNKTFQELAVFHDLEGMWEELSPKIWTFMENSQEMDLVRMLLDSRDNDHFWEQQLDGLDWTAQDIVAFL
AKHPEDVQSSNGSVYTWREAFNETNQAIRTISRFMECVNLNKLEPIATEVWLINKSMELLDERKFWAGIVFTGITPGSIELPHHVKYKIRMDIDNVERTNKIKDGYWDPGPRADPFEDMRYV
WGGFAYLQDVVEQAIIRVLTGTEKKTGVYMQQMPYPCYVDDIFLRVMSRSMPLFMTLAWIYSVAVIIKGIVYEKEARLKETMRIMGLDNSILWFSWFISSLIPLLVSAGLLVVILKLGNLLP
YSDPSVVFVFLSVFAVVTILQCFLISTLFSRANLAAACGGIIYFTLYLPYVLCVAWQDYVGFTLKIFASLLSPVAFGFGCEYFALFEEQGIGVQWDNLFESPVEEDGFNLTTSVSMMLFDTF
LYGVMTWYIEAVFPGQYGIPRPWYFPCTKSYWFGEESDEKSHPGSNQKRISEICMEEEPTHLKLGVSIQNLVKVYRDGMKVAVDGLALNFYEGQITSFLGHNGAGKTTTMSILTGLFPPTSG
TAYILGKDIRSEMSTIRQNLGVCPQHNVLFDMLTVEEHIWFYARLKGLSEKHVKAEMEQMALDVGLPSSKLKSKTSQLSGGMQRKLSVALAFVGGSKVVILDEPTAGVDPYSRRGIWELLLK
YRQGRTIILSTHHMDEADVLGDRIAIISHGKLCCVGSSLFLKNQLGTGYYLTLVKKDVESSLSSCRNSSSTVSYLKKEDSVSQSSSDAGLGSDHESDTLTIDVSAISNLIRKHVSEARLVED
IGHELTYVLPYEAAKEGAFVELFHEIDDRLSDLGISSYGISETTLEEIFLKVAEESGVDAETSDGTLPARRNRRAFGDKQSCLRPFTEDDAADPNDSDIDPESRETDLLSGMDGKGSYQVKG
WKLTQQQFVALLWKRLLIARRSRKGFFAQIVLPAVFVCIALVFSLIVPPFGKYPSLELQPWMYNEQYTFVSNDAPEDTGTLELLNALTKDPGFGTRCMEGNPIPDTPCQAGEEEWTTAPVPQ
TIMDLFQNGNWTMQNPSPACQCSSDKIKKMLPVCPPGAGGLPPPQRKQNTADILQDLTGRNISDYLVKTYVQIIAKSLKNKIWVNEFRYGGFSLGVSNTQALPPSQEVNDAIKQMKKHLKLA
KDSSADRFLNSLGRFMTGLDTKNNVKVWFNNKGWHAISSFLNVINNAILRANLQKGENPSHYGITAFNHPLNLTKQQLSEVALMTTSVDVLVSICVIFAMSFVPASFVVFLIQERVSKAKHL
QFISGVKPVIYWLSNFVWDMCNYVVPATLVIIIFICFQQKSYVSSTNLPVLALLLLLYGWSITPLMYPASFVFKIPSTAYVVLTSVNLFIGINGSVATFVLELFTDNKLNNINDILKSVFLI
FPHFCLGRGLIDMVKNQAMADALERFGENRFVSPLSWDLVGRNLFAMAVEGVVFFLITVLIQYRFFIRPRPVNAKLSPLNDEDEDVRRERQRILDGGGQNDILEIKELTKIYRRKRKPAVDR
ICVGIPPGECFGLLGVNGAGKSSTFKMLTGDTTVTRGDAFLNKNSILSNIHEVHQNMGYCPQFDAITELLTGREHVEFFALLRGVPEKEVGKVGEWAIRKLGLVKYGEKYAGNYSGGNKRKL
STAMALIGGPPVVFLDEPTTGMDPKARRFLWNCALSVVKEGRSVVLTSHSMEECEALCTRMAIMVNGRFRCLGSVQHLKNRFGDGYTIVVRIAGSNPDLKPVQDFFGLAFPGSVLKEKHRNM
LQYQLPSSLSSLARIFSILSQSKKRLHIEDYSVSQTTLDQVFVNFAKDQSDDDHLKDLSLHKNQTVVDVAVLTSFLQDEKVKESYVLEGSDEVDAVEGSHHHHHHHHHH
1|22|42|2|639|659|3|682|707|4|715|739|5|745|764|6|821|842|7|1344|1365|8|1654|1674|9|1703|1724|10|1732|1756|11|1765|1786|12
|1851|1867

5vai_R|461|7|4.1|E-MIC|0.54|298.57|37.39
MKTIIALSYIFCLVFADYKDDDDAAAGQPGNGSAFLLAPNRSHAPGGGGSLEVLFQGPGGGSRPQGATVSLSETVQKWREYRRQCQHFLTEAPPLATGLFCNRTFDDYACWPDGAPGSFVNV
SCPWYLPWASNVLQGHVYRFCTAEGHWLPKDNSSLPWRDLSECEESRRGEKSSPEERLLSLYIIYTVGYALSFSALVIASAILLGFRHLHCTRNYIHLNLFASFILRALSVFIKDAALKWMY
STAAQQHQWDGLLSYQDSLGCRLVFLLMQYCVAANYYWLLVEGAYLYTLLAFAVFSEQRIFKLYLSIGWGVPLLFVIPWGIVKYLYEDEGCWTRNSNMNYWLIIRLPILFAIGVNFLIFIRV
ICIVVSKLKANLMCKTDIKCRLAKSTLTLIPLLGTHEVIFAFVMDEHARGTLRFVKLFTELSFTSFQGLMVAILYCFVNNEVQMEFRKSWERWRL
1|144|164|2|179|200|3|224|249|4|268|290|5|304|329|6|350|372|7|379|404

5l25_A|652|14|4.11|X-RAY|0.53|429.65|81.1
MEETFVPFEGIKNDLKGRLMCYKQDWTGGFKAGFRILAPTTYIFFASAIPVISFGEQLERSTDGVLTAVQTLASTAICGMIHSIIGGQPLLILGVAEPTVIMYTFMFNFAKARPELGRDLFL
AWSGWVCVWTALMLFVLAICGACSIINRFTRVAGELFGLLIAMLFMQQAIKGLVDEFRIPERENQKLKEFLPSWRFANGMFALVLSFGLLLTGLRSRKARSWRYGTGWLRSLIADYGVPLMV
LVWTGVSYIPAGDVPKGIPRRLFSPNPWSPGAYGNWTVVKEMLDVPIVYIIGAFIPASMIAVLYYFDHSVASQLAQQKEFNLRKPSSYHYDLLLLGFLTLMCGLLGVPPSNGVIPQSPMHTK
SLATLKYQLLRNRLVATARRSIKTNASLGQLYDNMQEAYHHMQTPLVYQQPQGLKELKESTIQATTFTGNLNAPVDETLFDIEKEIDDLLPVEVKEQRVSNLLQSTMVGGCVAAMPILKMIP
TSVLWGYFAFMAIESLPGNQFWERILLLFTAPSRRFKVLEDYHATFVETVPFKTIAMFTLFQTTYLLICFGLTWIPIAGVMFPLMIMFLIPVRQYLLPRFFKGAHLQDLDAAEYEEAPALPF
NLAAETEIGSTTSYPGDLEILDEVMTRSRGEFRHTAAALVPR
1|36|63|2|68|81|3|97|113|4|120|139|5|158|178|6|198|220|7|239|252|8|292|313|9|335|348|10|358|368|11|465|483|12|489|504|13|5
50|563|14|564|578

5ldw_Y|141|4|4.27|E-MIC|0.52|91.62|16.21
MAKTVLRQYWDIPEGTECHRKTYATTSIGGAAGLVVSAYSVALKTPTSFLEGVARTGRYTFTAAAIGAIFGLTSCISAQVREKPDDPLNYLIGGCAGGLILGARTRSYGIGAAACAYMGLTA
ALVKMGQLEGWQVFAEPKV
1|20|42|2|48|76|3|86|103|4|109|124

4nv6_A|291|5|4.19|X-RAY|0.55|184.22|25.65
MASYLKLKAQEETWLQRHSRLILAILAGLGSLLTAYLTYTKLTEQPAAFCTGDGGCDLVLSSRWAEFLGIPTAAVGLLGFLGVLALAVLPDGLPLVKRWRWPALFGLVSAMTAFEMYMLYLM
VAVLRQFCMYCTTAIILVAGLGLVTVLGHRWLDGGKLAFSYILVAFLTLVTTIGVYANQVPPPSPLAVGLAAHLRQIGGTMYGAYWCPHAQDQKELFGAAFDQVPYVECSPNGPGTPQAQEC
TEAGITSYPTWIINGRTYTGVRSLEALAVASGYPLEEGRLEHHHHHH
1|20|38|2|72|88|3|99|120|4|131|150|5|157|178

5mkf_A|968|6|4.2|E-MIC|0.76|393.17|37.87
MVNSSRVQPQQPGDAKRPPAPRAPDPGRLMAGCAAVGASLAAPGGLCEQRGLEIEMQRIRQAAARDPPAGAAASPSPPLSSCSRQAWSRDNPGFEAEEEEEEVEGEEGGMVVEMDVEWRPGS
RRSAASSAVSSVGARSRGLGGYHGAGHPSGRRRRREDQGPPCPSPVGGGDPLHRHLPLEGQPPRVAWAERLVRGLRGLWGTRLMEESSTNREKYLKSVLRELVTYLLFLIVLCILTYGMMSS
NVYYYTRMMSQLFLDTPVSKTEKTNFKTLSSMEDFWKFTEGSLLDGLYWKMQPSNQTEADNRSFIFYENLLLGVPRIRQLRVRNGSCSIPQDLRDEIKECYDVYSVSSEDRAPFGPRNGTAW
IYTSEKDLNGSSHWGIIATYSGAGYYLDLSRTREETAAQVASLKKNVWLDRGTRATFIDFSVYNANINLFCVVRLLVEFPATGGVIPSWQFQPLKLIRYVTTFDFFLAACEIIFCFFIFYYV
VEEILEIRIHKLHYFRSFWNCLDVVIVVLSVVAIGINIYRTSNVEVLLQFLEDQNTFPNFEHLAYWQIQFNNIAAVTVFFVWIKLFKFINFNRTMSQLSTTMSRCAKDLFGFAIMFFIIFLA
YAQLAYLVFGTQVDDFSTFQECIFTQFRIILGDINFAEIEEANRVLGPIYFTTFVFFMFFILLNMFLAIINDTYSEVKSDLAQQKAEMELSDLIRKGYHKALVKLKLKKNTVDDISESLRQG
GGKLNFDELRQDLKGKGHTDAEIEAIFTKYDQDGDGQELTEHEHQQMRDDLEKEREDLDLDHSSLPRPMSSRSFPRSLDDSEEDDDEDSGHSSRRRGSISSGVSYEEFQVLVRRVDRMEHSIG
SIVSKIDAVIVKLEIMERAKLKRREVLGRLLDGVAEDERLGRDSEIHREQMERLVREELERWESDDAASQISHGLGTPVGLNGQPRPRSSRPSSSQSTEGMEGAGGNGSSNVHV
1|221|241|2|471|493|3|506|528|4|556|578|5|597|620|6|654|677

5ldw_a|70|1|4.27|E-MIC|0.56|42.03|3.87
MWFEVLPGIAVMGVCLFIPGMATARIHRFSNGGKEKRVAHYPYQWYLMERDRRVSGVNRSYVSKGLENID
1|5|27

5ldw_c|49|1|4.27|E-MIC|0.53|30.57|3.43
KFYIQEPPHGSPNWLKVGLTLGTSVFLWIYLIKQHNEDVLEYKRRNGLE
1|14|32

5ldw_o|136|2|4.27|E-MIC|0.89|28.49|9.03
GAHLARRYLGDASVEPDPLRMPTFPPDYGFPERKEREMVATQQEMNDAQLVLQQRDYCAHYLIRFLKCKRDSFPNFLACKHERHDWDYCEHLDYVKRMKEFERERRLLQRKKRREQREADMA
KGLGPGEVAPEVAL
1|72|75|2|77|80

4hw9_A|309|2|4.14|X-RAY|0.81|102.37|11.25
MGSSHHHHHHSSGLVPRGSHTLINMDEIKTLLVDFFPQAKHFGIILIKAVIVFCIGFYFSFFLRNKTMKLLSKKDEILANFVAQVTFILILIITTIIALSTLGVQTTSIITVLGTVGIAVAL
ALKDYLSSIAGGIILIILHPFKKGDIIEISGLEGKVEALNFFNTSLRLHDGRLAVLPNRSVANSNIINSNNTACRRIEWVCGVGYGSDIELVHKTIKDVIDTMEKIDKNMPTFIGITDFGSS
SLNFTIRVWAKIEDGIFNVRSELIERIKNALDANHIEIPFNKLDIAIKNQDSPKLINDYKDDDDK
1|20|39|2|66|78

5lc5_q|138|1|4.35|E-MIC|0.88|31.28|4.87
*ELLQVLKRGLQQVSGHGGLRGYLRVLFRANDVRVGTLVGEDKYGNKYYEDNKQFFGRHRWVIYTTEMNGKNTFWDVDGSMVPPEWHRWLHCMTDDPPTVKPPTARKFIWTNHKFNLSGTPQ
QYVPYSTTRKKIQEWVP
1|4|12

4o9u_D|450|9|6.93|X-RAY|0.52|300.77|47.97
MDLIQAAYFVVAILFIVGLKRMAHPTTAKSGIVWAGWGMVLAVLATFFWPGMGNFALILLALLLGSVVAWWAAVRVAMTDMPQMVAIYNGMGGGAAATIAAVELLKGAFENTGLMALAILGG
LIGSVAFTGSLIAFAKLQGIMKSRPILFPGQKAVNALVLALTVVIGLSLLWNDATASIVLFFLLALLFGVLMTLPIGGGDMPVAISFYNAFTGMAVGFEGFAVGNPALMVAGTLVGAAGTLL
TVLMARAMNRSVWSVLVGGFGVEQEAGEVKGSLKPIDVEDAAVMLAYAGKVVFVPGYGMALSQAQHKLKELADLLEARGVEVKFAIHPVAGRMPGHMNVLLAEAGVDYDKLKDLEEINPEFP
TVDVAVVIGANDVVNPAARRPGSPLYGMPILDVDKAKNVIVIKRGQGKGFAGVENELFYAENTRMLYGDAQKVLTELIQALKRL
1|3|20|2|31|48|3|55|73|4|84|105|5|112|138|6|153|172|7|177|198|8|206|227|9|228|242

3din_E|76|2|4.5|X-RAY|0.5|50.26|7.3
MKTFFLIVHTIISVALIYMVQVQMSKFSELGGAFGSGGLHTVFGRRKGLDTGGKITLVLSVLFFVSCVVTAFVLTR
1|11|30|2|53|71

3j2w_Q|81|1|5.0|E-MIC|0.6|45.06|4.09
****
****
****
****
****
****
**********STNGTAHGHPHE
IILYYYELYPTMTVVIVSVASFVLLSMVGTAVGMCVCARRRCITPYELTPGATVPFLLSLLCCVRTTKA
1|871|892

3fh6_F|480|7|4.5|X-RAY|0.59|282.91|40.27
**********MAQGEYLFAITTLILSSAGLYIFANRKAYAWRYVYPGMAGMGLFVLFPLVCTIAIAFTNYSSTNQLTFERAQEVLLDRSWQAGKTYNF
GLYPAGDEWQLALSDGETGKNYLSDAFKFGGEQKLQLKETTAQPEGERANLRVITQNRQALSDITAILPDGNKVMMSSLRQFSGTQPLYTLDGDGTLTNNQSGVKYRPNNQIGFYQSITADG
NWGDEKLSPGYTVTTGWKNFTRVFTDEGIQKPFLAIFVWTVVFSLITVFLTVAVGMVLACLVQWEALRGKAVYRVLLILPYAVPSFISILIFKGLFNQSFGEINMMLSALFGVKPAWFSDPT
TARTMLIIVNTWLGYPYMMILCMGLLKAIPDDLYEASAMDGAGPFQNFFKITLPLLIKPLTPLMIASFAFNFNNFVLIQLLTNGGPDRLGTTTPAGYTDLLVNYTYRIAFEGGGGQDFGLAA
AIATLIFLLVGALAIVNLKATRMKFD
1|42|56|2|68|89|3|277|306|4|314|326|5|369|389|6|426|439|7|487|507

5wua_E|1582|17|5.6|E-MIC|0.65|inf|110.46
MPLAFCGTENHSAAYRVDQGVLNNGCFVDALNVVPHVFLLFITFPILFIGWGSQSSKVHIHHSTWLHFPGHNLRWILTFILLFVLVCEIAEGILSDGVTESRHLHLYMPAGMAFMAAITSVV
YYHNIETSNFPKLLIALLIYWTLAFITKTIKFVKFYDHAIGFSQLRFCLTGLLVILYGMLLLVEVNVIRVRRYIFFKTPREVKPPEDLQDLGVRFLQPFVNLLSKGTYWWMNAFIKTAHKKP
IDLRAIGKLPIAMRALTNYQRLCVAFDAQARKDTQSPQGARAIWRALCHAFGRRLILSSTFRILADLLGFAGPLCIFGIVDHLGKENHVFQPKTQFLGVYFVSSQEFLGNAYVLAVLLFLAL
LLQRTFLQASYYVAIETGINLRGAIQTKIYNKIMHLSTSNLSMGEMTAGQICNLVAIDTNQLMWFFFLCPNLWAMPVQIIVGVILLYYILGVSALIGAAVIILLAPVQYFVATKLSQAQRST
LEHSNERLKQTNEMLRGMKLLKLYAWESIFCSRVEVTRRKEMTSLRAFAVYTSISIFMNTAIPIAAVLITFVGHVSFFKESDLSPSVAFASLSLFHILVTPLFLLSSVVRSTVKALVSVKKL
SEFLSSAEIREEQCAPREPAPQGQAGKYQAVPLKVVNRKRPAREEVRDLLGPLQRLAPSMDGDADNFCVQIIGGFFTWTPDGIPTLSNITIRIPRGQLTMIVGQVGCGKSSLLLATLGEMQK
VSGAVFWNSNLPDSEGEDPSSPERETAAGSDIRSRGPVAYASQKPWLLNATVEENITFESPFNKQRYKMVIEACSLQPDIDILPHGDQTQIGERGINLSGGQRQRISVARALYQQTNVVFLD
DPFSALDVHLSDHLMQAGILELLRDDKRTVVLVTHKLQYLPHADWIIAMKDGTIQREGTLKDFQRSECQLFEHWKTLMNRQDQELEKETVMERKASEPSQGLPRAMSSRDGLLLDEEEEEEE
AAESEEDDNLSSVLHQRAKIPWRACTKYLSSAGILLLSLLVFSQLLKHMVLVAIDYWLAKWTDSALVLSPAARNCSLSQECDLDQSVYAMVFTLLCSLGIVLCLVTSVTVEWTGLKVAKRLH
RSLLNRIILAPMRFFETTPLGSILNRFSSDCNTIDQHIPSTLECLSRSTLLCVSALTVISYVTPVFLVALLPLAVVCYFIQKYFRVASRDLQQLDDTTQLPLLSHFAETVEGLTTIRAFRYE
ARFQQKLLEYTDSNNIASLFLTAANRWLEVRMEYIGACVVLIAAATSISNSLHRELSAGLVGLGLTYALMVSNYLNWMVRNLADMEIQLGAVKRIHALLKTEAESYEGLLAPSLIPKNWPDQ
GKIQIQNLSVRYDSSLKPVLKHVNALISPGQKIGICGRTGSGKSSFSLAFFRMVDMFEGRIIIDGIDIAKLPLHTLRSRLSIILQDPVLFSGTIRFNLDPEKKCSDSTLWEALEIAQLKLVV
KALPGGLDAIITEGGENFSQGQRQLFCLARAFVRKTSIFIMDEATASIDMATENILQKVVMTAFADRTVVTIAHRVHTILSADLVMVLKRGAILEFDKPETLLSQKDSVFASFVRADK
1|30|52|2|71|92|3|107|127|4|135|155|5|165|188|6|302|324|7|355|376|8|436|451|9|458|477|10|541|562|11|573|594|12|1009|1032|1
3|1064|1086|14|1140|1160|15|1162|1179|16|1248|1268|17|1279|1299

4v6m_AZ|98|1|7.1|E-MIC|0.66|48.59|4.36
GTRLAGILFLLTVLTTVLVSGWVVLGWMEDAQRLPLSKLVLTGERHYTRNDDIRQSILALGEPGTFMTQDVNIIQTQIEQRLQHARLDKPGARHPCWP**********
****
1|25|45

5kk2_A|889|4|7.3|E-MIC|0.82|283.91|26.76
MQKIMHISVLLSPVLWGLIFGVSSNSIQIGGLFPRGADQEYSAFRVGMVQFSTSEFRLTPHIDNLEVANSFAVTNAFCSQFSRGVYAIFGFYDKKSVNTITSFCGTLHVSFITPSFPTDGTH
PFVIQMRPDLKGALLSLIEYYQWDKFAYLYDSDRGLSTLQAVLDSAAEKKWQVTAINVGNINNDKKDETYRSLFQDLELKKERRVILDCERDKVNDIVDQVITIGKHVKGYHYIIANLGFTD
GDLLKIQFGGANVSGFQIVDYDDSLVSKFIERWSTLEEKEYPGAHTATIKYTSALTYDAVQVMTEAFRNLRKQRIEISRRGNAGDCLANPAVPWGQGVEIERALKQVQVEGLSGNIKFDQNG
KRINYTINIMELKTNGPRKIGYWSEVDKMVVTLTELPSGNDTSGLENKTVVVTTILESPYVMMKKNHEMLEGNERYEGYCVDLAAEIAKHCGFKYKLTIVGDGKYGARDADTKIWNGMVGEL
VYGKADIAIAPLTITLVREEVIDFSKPFMSLGISIMIKKPQKSKPGVFSFLDPLAYEIWMCIVFAYIGVSVVLFLVSRFSPYEWHTEEFEDGRETQSSESTNEFGIFNSLWFSLGAFMRQGC
DISPRSLSGRIVGGVWWFFTLIIISSYTANLAAFLTVERMVSPIESAEDLSKQTEIAYGTLDSGSTKEFFRRSKIAVFDKMWTYMRSAEPSVFVRTTAEGVARVRKSKGKYAYLLESTMNEY
IEQRKPCDTMKVGGNLDSKGYGIATPKGSSLGNAVNLAVLKLNEQGLLDKLKNKWWYDKGECGSGGGDSKEKTSALSLSNVAGVFYILVGGLGLAMLVALIEFCYKSRAEAKRMKVAKNPQN
INPSSSQNSQNFATDYKDDDDKEGYNVYGIESVKI
1|521|545|2|571|585|3|596|617|4|791|816

5kk2_F|323|4|7.3|E-MIC|0.6|188.1|21.85
MGLFDRGVQMLLTTVGAFAAFSLMTIAVGTDYWLYSRGVCKTKSVSENETSKKNEEVMTHSGLWRTCCLEGNFKGLCKQIDHFPEDADYEADTAEYFLRAVRASSIFPILSVILLFMGGLCI
AASEFYKTRHNIILSAGIFFVSAGLSNIIGIIVYISANAGDPSKSDSKKNSYSYGWSFYFGALSFIIAEMVGVLAVHMFIDRHKQLRATARATDYLQASAITRIPSYRYRYQRRSRSSSRST
EPSHSRDASPVGVKGFNTLPSTEISMYTLSRDPLKAATTPTATYNSDRDNSFLQVHNCIQKDSKDSLHANTANRRTTPV
1|10|28|2|100|123|3|134|158|4|178|199

4aq9_C|522|4|6.2|E-MIC|0.71|241.52|23.58
MGNIHFVYLLISCLYYSGCSGVNEEERLINDLLIVNKYNKHVRPVKHNNEVVNIALSLTLSNLISLKETDETLTTNVWMDHAWYDHRLTWNASEYSDISILRLRPELIWIPDIVLQNNNDGQ
YNVAYFCNVLVRPNGYVTWLPPAIFRSSCPINVLYFPFDWQNCSLKFTALNYNANEISMDLMTDTIDGKDYPIEWIIIDPEAFTENGEWEIIHKPAKKNIYGDKFPNGTNYQDVTFYLIIRR
KPLFYVINFITPCVLISFLAALAFYLPAESGEKMSTAICVLLAQAVFLLLTSQRLPETALAVPLIGKYLMFIMSLVTGVVVNCGIVLNFHFRTPSTHVLSTRVKQIFLEKLPRILHMSRVDE
IEQPDWQNDLKLRRSSSVGYISKAQEYFNIKSRSELMFEKQSERHGLVPRVTPRIGFGNNNENIAASDQLHDEIKSGIDSTNYIVKQIKEKNAYDEEVGNWNLVGQTIDRLSMFIITPVMVL
GTIFIFVMGNFNRPPAKPFEGDPFDYSSDHPRCA
1|229|250|2|256|278|3|292|314|4|456|480

4cg6_D|17|1|7.8|E-MIC|0.79|4.91|1.1
VFIVSVGSFISVLFIVI
1|3|17

2ww9_B|80|1|8.6|E-MIC|0.56|48.09|4.01
MARASEKGEEKKQSNNQVEKLVEAPVEFVREGTQFLAKCKKPDLKEYTKIVKAVGIGFIAVGIIGYAIKLIHIPIRYVIV
1|47|72

4v7i_BD|206|1|9.6|E-MIC|0.79|73.68|5.2
MARYLGPKLKLSRREGTDLFLKSGVRAIDTKCKIEQAPGQHGARKPRLSDYGVQLREKQKVRRIYGVLERQFRNYYKEAARLKGNTGENLLALLEGRLDNVVYRMGFGATRAEARQLVSHKA
IMVNGRVVNIASYQVSPNDVVSIREKAKKQSRVKAALELAEQREKPTWLEVDAGKMEGTFKRKPERSDLSADINEHLIVELYSK
1|40|64

4v7i_BE|167|1|9.6|E-MIC|0.81|54.78|5.01
MAHIEKQAGELQEKLIAVNRVSKTVKGGRIFSFTALTVVGDGNGRVGFGYGKAREVPAAIQKAMEKARRNMINVALNNGTLQHPVKGVHTGSRVFMQPASEGTGIIAGGAMRAVLEVAGVHN
VLAKAYGSTNPINVVRATIDGLENMNSPEMVAAKRGKSVEEILGK
1|33|50

4v7i_BC|233|10|9.6|E-MIC|0.71|105.72|36.99
MGQKVHPNGIRLGIVKPWNSTWFANTKEFADNLDSDFKVRQYLTKELAKASVSRIVIERPAKSIRVTIHTARPGIVIGKKGEDVEKLRKVVADIAGVPAQINIAEVRKPELDAKLVADSITS
QLERRVMFRRAMKRAVQNAMRLGAKGIKVEVSGRLGGAEIARTEWYREGRVPLHTLRADIDYNTSEAHTTYGVIGVKVWIFKGEILGGMAAVEQPEKPAAQPKKQQRKGRK
1|31|42|2|70|88|3|110|132|4|137|158|5|169|192|6|207|226|7|263|284|8|320|338|9|391|407|10|408|421

2k3m_A|151|1|N-APP|S-NMR|0.74|62.03|4.86
MHHHHHHSSGVDLGTENLYFQSNAMSDFDTERVSRAVAAALVGPGGVALVVKVCAGLPGVIHTPARRGFFRCNPERIQIGDWRYEVAHDGRLLAAHMVNGIVIAEDALIAEAVGPHLARALG
QIVCRYGATVIPNINAAIEVLGTGTDYRF
1|8|30

2lom_A|93|2|N-APP|S-NMR|0.54|58.04|8.24
MSTDTGVSLPSYEEDQGSKLIRKAKEAPFVPVGIAGFAAIVAYGLYKLKSRGNTKMSIHLIHMRVAAQGFVVGAMTVGMGYSMYREFWAKPKP
1|29|44|2|65|81

2mj2_A|36|1|N-APP|S-NMR|0.5|22.93|2.94
TWSGTKKRAQRILIFLLEFLLDFCTGEDSVDGKKRQ
1|7|24

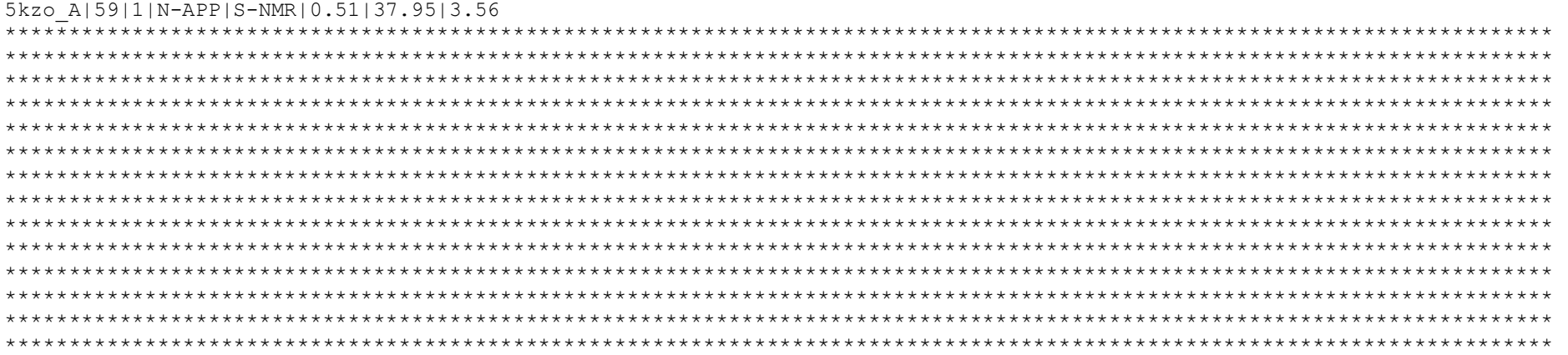
2l9u_A|40|1|N-APP|S-NMR|0.53|24.42|2.77
****
****
****
****
****
**********MGRTHLTMALTVIAGLVVIFMMLGGTFLYWRGRRHHHHHH
1|641|665

2ksr_A|164|4|N-APP|S-NMR|0.5|110.89|17.01
MHHHHHHSTSVDLGTENLYFQSNARRKPLFYTINLIIPCVLITSLAILVFYLPSDCGEKMTLCISVLLALTVFLLLISKIVPPTSLDVPLVGKYLMFTMVLVTFSIVTSVCVLNVHHRSPTT
HTPRGGGGYVAMVIDRLFLWIFVFVCVFGTIGMFLQPLFQNY
1|29|51|2|57|77|3|94|111|4|137|157

2lck_A|303|6|N-APP|S-NMR|0.5|206.66|30.51
**********MTVKFLGAGTAACIADLITFPLDTAKVRLQIQGESQGLVRTAASAQYRGVLGTILTMVRTEGPRSLYNGLVAGLQRQMSFASVRIGLYDSVKQFYTKGSEHAGIGSRLLA
GSTTGALAVAVAQPTDVVKVRFQAQARAGGGRRYQSTVEAYKTIAREEGIRGLWKGTSPNVARNAIVNCAELVTYDLIKDTLLKANLMTDDLPCHFTSAFGAGFCTTVIASPVDVVKTRYMN
SALGQYHSAGHCALTMLRKEGPRAFYKGFMPSFLRLGSWNVVMFVTYEQLKRALMAAYQSREAPFHHHHHH
1|17|38|2|75|104|3|116|142|4|175|200|5|222|242|6|275|293

2m3b_A|52|1|N-APP|S-NMR|0.51|33.23|3.43
MEKVQYLTRSAIRRASTIEMPQQARQNLQNLFINFCLILICLLLICIIVMLL
1|29|50

2lor_A|108|2|N-APP|S-NMR|0.56|65.53|8.62
MVNLGLSRVDDAVAAKHPGLGEYAACQSHAFMKGVFTFVTGTGMAFGLQMFIQRKFPYPLQWSLLVAVVAGSVVSYGVTRVESEKCNNLWLFLETGQLPKDRSTDQRS
1|30|47|2|62|78

5kzo_A|59|1|N-APP|S-NMR|0.51|37.95|3.56

****MGHHHHHHVQSETVEPPPPAQLHFMYVAAAAFVLLFFVGCGVLLSRKRRRQHGQLWFPE
1|1732|1756

3zd0_A|86|2|N-APP|S-NMR|0.5|57.15|7.59
GPLGSPEFAAMDYKDDDDKALENLVVLNAASVAGAHGILSFLVFFCAAWYIKGRLAPGAAYAFYGVWPLLLLLLALPPRAYAAAAS
1|24|46|2|56|75

2m7g_A|61|1|N-APP|S-NMR|0.58|34.81|3.78
FTLIELLIVVAIIGILAAIAIPQFSAYRVKAYNSAASSDLRNLKTALESAFADDQTYPPES
1|4|21

2k74_A|183|4|N-APP|S-NMR|0.52|121.04|18.13
MLRFLNQASQGRGAWLLMAFTALALELTALWFQHVMLLKPCVLSIYERAALFGVLGAALIGAIAPKTPLRYVAMVIWLYSAFRGVQLTYEHTMLQLYPSPFATSDFMVRFPEWLPLDKWVPQ
VFVASGDCAERQWDFLGLEMPQWLLGIFIAYLIVAVLVVISQPFKAKKRDLFGRGHHHHHH
1|12|32|2|47|62|3|68|86|4|144|162

2m20_A|60|1|N-APP|S-NMR|0.52|38.12|3.61
KIPSIATGLVGALLLLLVVALGIGLFIRRRHIVRKRTLRRLLQERELVEPLTPSGEKLWS
1|4|27

2jo1_A|72|1|N-APP|S-NMR|0.57|42.81|3.91
ESPKEHDPFTYDYQSLQIGGLVIAGILFILGILIVLSRRCRCKFNQQQRTGEPDEEEGTFRSSIRRLSTRRR
1|15|37

2kse_A|186|2|N-APP|S-NMR|0.66|94.18|9.99
MGKFTQRLSLRVRLTLIFLILASVTWLLSSFVAWKQTTDNVDELFDTQLMLFAKRLSTLDLNEINAADRMAQTPNRLKHGHVDDDALTFAIFTHDGRMVLNDGDNGEDIPYSYQREGFADGQ
LVGEDDPWRFVWMTSPDGKYRIVVGQEWEYREDMALAIVAGQLIPWLVALPIMLIIMMVLLGRE
1|13|31|2|161|181

2lp1_A|122|1|N-APP|S-NMR|0.73|52.1|4.63
****
****
****
****
****
**MDAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIATVIVITLVMLKKKQYTSIH
HGVVEVDAAVTPEERHLSKMQQNGYENPTYKFFEQMQNQGRILQISITLAAALEHHHHHH
1|704|723