UNIVERZITET U BEOGRADU

MATEMATIČKI FAKULTET

Vladimir R. Perović

Razvoj multifunkcionalne bioinformatičke platforme zasnovane na potencijalu elektron-jon interakcije bioloških molekula

doktorska disertacija

Beograd, 2013

UNIVERSITY OF BELGRADE

FACULTY OF MATHEMATICS

Vladimir R. Perović

Development of multifunctional bioinformatics platform based on electron-ion interaction potential of biological molecules

Doctoral Dissertation

Belgrade, 2013

Mentor

Prof. dr Dušan D. Tošić,
redovni profesor Matematičkog fakulteta Univerziteta u Beogradu

Članovi komisije

Dr Veljko J. Veljković,
naučni savetnik Instituta za nuklearne nauke „Vinča“

Prof. dr Miodrag V. Živković,
redovni profesor Matematičkog fakulteta Univerziteta u Beogradu

Datum odbrane

_____

Predgovor

Metoda informacionih spektara, koja je 1983. godine razvijena u Institutu VINČA, danas se koristi u preko 70 laboratorija širom sveta. Rezultati dobijeni primenom ove bioinformatičke metode do sada su objavljeni u preko 100 naučnih radova i nekoliko monografija, a poslužili su i kao osnova za više međunarodnih patenata. ISM je takođe uključen u bioinformatičke kurseve koji se drže na univerzitetima u Engleskoj, Hrvatskoj, SAD, Kini, Indiji, Japanu, Brazilu, Australiji i Kanadi.

U ovoj disertaciji je prikazan doprinos u razvoju EIIP/ISM bioinformatičke platforme koji se ogleda u nizu originalnih softverskih programa koji predstavljaju osnovu ključnih modula ove platforme. Pored toga, razvijeni su novi metodološki pristupi koji su ugrađeni u EIIP/ISM bioinformatičku platformu, od kojih je svakako najznačajniji onaj koji se odnosi na funkcionalnu filogenetsku analizu zasnovanu na informacionim spektrima proteina. Rezultati dobijeni primenom EIIP/ISM bioinformatičke platforme u izučavanju različitih bioloških fenomena i bioloških sistema (evolucija virusa gripa, molekularni mehanizmi odbrane od malignih bolesti zasnovani na fizičkoj aktivnosti, selekcija kandidata za lekove protiv SIDE, uticaj mutacija na razvoj kardiovaskularnih obolenja, itd.) objavljeni su u 13 radova (6 u vrhunskim međunarodnim časopisma (M21) i 5 u istaknutim međunarodnim časopisma (M22)), na kojima sam učestvovao kao koautor i bio prvi autor u dva rada.

Na kraju treba istaći da su moduli EIIP/ISM bioinformatičke platforme (Osnovni ISM modul, Modul za procenu biološkog efekta mutacija, Modul za modifikaciju proteina i dizajniranje peptida i Modul za pretraživanje molekulskih biblioteka), predstavljali osnovni bioinformatički alat u okviru evropskog FP6 projekta „Targeting replication and integration of HIV“ (TRIOH). Jedan od najvažnijih rezultata ovog projekta, u čijoj realizaciji je učestvovalo 32 instituta i univerziteta iz 11 evropskih zemalja, predstavlja novi inhibitor HIV integraze čije se otkriće zasnivalo na neposrednoj primeni EIIP/ISM bioinformatičke platforme.

Zahvalnica

Ova doktorska disertacija je osmišljena i urađena u Laboratoriji za multidisciplinarna istraživanja Instituta za nuklearne nauke “Vinča“.

Najveću zahvalnost dugujem profesoru dr Dušanu Tošiću, mom mentoru, na stručnoj pomoći, korisnim savetima i podršci tokom izrade disertacije. Takođe, neizmerno sam zahvalan dr Veljku Veljkoviću na ukazanom poverenju, razumevanju, strpljenju, podsticaju, ohrabrenju, kao i na nesebičnoj podršci i pomoći oko izrade ove teze. Zahvaljujem se i profesoru dr Miodragu Živkoviću, članu komisije, na stručnim i prijateljskim savetima.

Želim da izrazim posebnu zahvalnost dr Sanji Glišić (viši naučni saradnik INN Vinča) na nesebičnoj pomoći, stručnim i prijateljskim savetima, kao i dr Neveni Veljković (naučni savetnik INN “Vinča“) na korisnim i stručnim savetima tokom izrade ove disertacije.

Svim kolegama iz moje laboratorije se zahvaljujem na svakodnevnoj podršci i prijateljskoj atmosferi. Posebno, hvala Branislavi Gemović i Nebojši Škrbiću na korisnim, praktičnim i originalnim savetima i pomoći prilikom izrade teze.

Na kraju, želim da se zahvalim svojoj porodici na ljubavi i velikoj podršci, a posebno roditeljima na usmeravanju mojih stremljenja ka obrazovanju i nauci. Takođe, hvala i najbližim prijateljima na entuzijazmu, podsticanju da istrajem, pravom prijateljstvu, pozitivnim savetima i ohrabrenju.

Razvoj multifunkcionalne bioinformatičke platforme zasnovane na potencijalu elektron-jon interakcije bioloških molekula

Rezime

Iako dalekodosežne međumolekulske interakcije (interakcije na rastojanjima >5Å) igraju značajnu ulogu u prepoznavanju i selektivnom međusobnom privlačenju molekula u biološkim sistemima, do sada nije razvijen odgovarajući softverski paket koji bi omogućio da se ova važna osobina uključi u izučavanje biološki aktivnih molekula. U ovom radu je razvijena multifunkcionalna softverska EIIP/ISM platforma zasnovana na fizičkim parametrima organskih molekula koji definišu njihovu dalekodosežnu interakciju. Ova platforma omogućava (i) izučavanje protein-protein interakcije i interakcije između proteina i malih molekula, (ii) izučavanje veze između strukture i funkcije proteina, (iii) procenu uticaja mutacija na biološku funkciju proteina, (iv) praćenje funkcionalne evolucije proteina, (v) dizajniranje molekula željene biološke aktivnosti i (vi) selekciju potencijalnih terapeutskih molekula. Rezultati primene EIIP/ISM platforme na različite konkretne probleme kao što su evolucija virusa gripa, analiza mutacija na LPL proteinu koje predstavljaju faktor rizika za nastanak kardiovaskularnih bolesti, identifikacija terapeutskih targeta za viruse HIV-1 i viruse gripa, selekcija kandidata za antibiotike i lekove za SIDU virtuelnim pretraživanjem molekulskih biblioteka, koji su prikazani u ovom radu, potvrdili su primenjivost ove platforme u rešavanju širokog spektra problema u molekularnoj biologiji, biomedicini i farmaciji.

Ključne reči: softverski paket, metoda informacionih spektara, Furijeova transformacija, filogenetska analiza, potencijal elektron-jon interakcije, proteinske sekvence, influenca virus, otkrivanje lekova

Naučna oblast: Informatika
Uža naučna oblast: Bioinformatika
UDK broj: 004.415:[004.6:[577.2+577+616.921.5]](043.3)

Development of multifunctional bioinformatics platform based on electron-ion interaction potential of biological molecules

Abstract

Although long-range intermolecular interactions (interactions acting on distances >5Å) play an important role in recognition and targeting between molecules in biological systems, there is no one appropriate software package allowing use of this important property in investigation of biologically active molecules. The multifunctional EIIP/ISM software, which is based on physical parameters determining long-range molecular properties, was developed in this thesis. This novel and unique platform allows (i) investigation of protein-protein and protein-small molecule interactions, (ii) analysis of structure/function relationship of proteins, (iii) assessment of biological effects of mutations in proteins, (iv) monitoring of the functional evolution of proteins, (v) "de novo" design of molecules with desired biological function and (vi) selection of candidate therapeutic molecules. Results of application of the EIIP/ISM platform on diverse problems (e.g. the evolution of influenza A viruses, assessment of biological effects of mutations on the LPL protein, representing a risk factor for cardiovascular diseases, identification of therapeutic targets for HIV and influenza viruses, virtual screening of molecular libraries for candidate antibiotics and anti-HIV drugs) which are presented in this thesis, confirm the applicability of this platform on broad spectrum of problems in molecular biology, biomedicine and pharmacology.

Keywords: Software package, Informational spectrum method, Fourier transform, Phylogenetic analysis, Electron-ion interaction potential, Protein sequence, Influenza virus, Drug discovery

Scientific field: Computer science

Scientific subfield: Bioinformatics

UDK number: 004.415:[004.6:[577.2+577+616.921.5]](043.3)

Skraćenice

AQVN: srednji kvazivalentni broj (eng. *average quasivalence number*)

BISTIC: konzorcijum biomedicinsko informatičkih nauka i tehnologija (eng. *Biomedical Information Science and Technology Initiative Consortium*)

ChemDB: baza malih molekula, koja sadrži testirane terapeutike za HIV, tuberkulozu i ostale oportunističke infekcije, čiji je vlasnik NIAID

CIS: konsenzus informacioni spektar (eng. *consensus informational spectrum*)

CS: krosspektar (eng. *cross-spectrum*)

DDBJ: Japanska baza DNK sekvenci (eng. *DNA Data Bank of Japan*)

DFT: diskretna Furijeova transformacija (eng. *discrete Fourier transform*)

DIT: razbijanje po vremenu (eng. *decimation in time*)

DNK: dezoksiribonukleinska kiselina (eng. *deoxyribonucleic acid, DNA*)

EBI: Evropski bioinformatički institut (eng. *European Bioinformatics Institute*)

EIIP: potencijal elektron-jon interakcije (eng. *electron-ion interaction potential*)

EMBL: Evropska biblioteka nukleotidnih sekvenci (eng. *European Molecular Biology Laboratory*)

FFT: brza Furijeova transformacija (eng. *fast Fourier transform*)

GenBank: baza nukleotidnih sekvenci, čiji je vlasnik NCBI

HIV: virus humane imunodeficijencije (eng. *human immunodeficiency virus*)

HPAIV: visoko patogeni ptičiji influenca virus (eng. *highly pathogenic avian influenza virus*)

HTS: visokopropusni skrining (eng. *high-throughput screening*)

IDFT: inverzna diskretna Furijeova transformacija (eng. *inverse discrete Fourier transform*)

IS: informacioni spektar (eng. *informational spectrum*)

ISM: metoda informacionih spektara (eng. *informational spectrum method*)

ISTREE: filogenetsko stablo zasnovana na informacionom spektru (eng. *informational spectrum-based phylogenetic tree*).

JTT: supstitucioni model za definisanje rastojanja između proteinskih sekvenci (eng. *Jones-Taylor-Thornton*)

LPL: lipoproteinska lipaza (eng. *lipoprotein lipase*)

ML: maksimalna verodostojnost (eng. *maximum likelihood*)

MP: maksimalna parsimonija (eng. *maximum parsimony*)

MSA: višestruko poravnavanje sekvenci (eng. *multiple* s*equence alignment*)

NCBI: Američki nacionalni centar za biotehnološke informacije (eng. *National Center for Biotechnology Information*)

NIAID: Američki nacionalni institut za alergije i infektivne bolesti (eng. *National Institute of Allergy and Infectious Diseases*)

NIH: Američki nacionalni instituti za zdravlje (eng. *National Institutes of Health*)

NJ: metoda spajanja suseda (eng. *neighbor-joining*) u hijerarhijskom klasterisanju

PDB: banka podataka o proteinima (eng. *Protein Data Bank*)

PubChem: Američka baza informacija o biološkim aktivnostima malih molekula

QSAR: kvantitativni odnos strukture i dejstva aktivnih supstanci (eng. *quantitative structure-activity relationship*)

RNK: ribonukleinska kiselina (eng. *ribonucleic acid, RNA*)

S/N: odnos vrednosti signal/šum (eng. *signal/noise*)

SIB: Švajcarski instituta za bioinformatiku (eng. *Swiss Institute of Bioinformatics*)

SMILES: pojednostavljena specifikacija linijskog unosa podataka o strukturi molekula (eng. *simplified molecular-input line-entry system*)

TrEMBL - baza prevedenih EMBL nukleotidnih sekvenci (eng. *Translated EMBL Nucleotide Sequence Data Library*)

UniProt: univerzalni proteinski resursi (eng. *universal protein resource*)

UPGMA: metoda hijerarhijskog klasterisanja zasnovan na aritmetičkoj sredini (eng. *Unweighted Pair Group Method with Arithmetic Mean*)

VS: virtuelni skrining (eng. *virtual screening*)

Sadržaj

1. Uvod

1.1. Bioinformatika

1.1.1. Opšte o bioinformatici

Biološki podaci koji se generišu, rastu velikom brzinom. Od 1979. do 2013. godine, broj nukleotidnih sekvenci u bazi GenBank porastao je od 2 hiljade do preko 162 miliona sekvenci [1], dok je Swiss-Prot baza proteinskih sekvenci od osnivanja 1986. godine dostigla broj od preko 540 hiljada sekvenci u 2013. godini [2]. Pored toga, broj hemijskih jedinjenja u bazi PubChem je od 2004. godine kada je osnovana, do 2013. godine dostigao broj od 119 miliona hemijskih jedinjenja [3]. Zajedno sa velikim brojem podataka o strukturi proteina, ekspresiji gena, interakciji proteina, nastala je ogromna količina i velika raznovrsnost bioloških informacija. Kao rezultat toga, računari su postali neophodni u biološkim istraživanjima.

Bioinformatika primenjuje računarske tehnike za razumevanje i organizaciju informacija o biološkim makromolekulima. Postoje tri osnovna cilja bioinformatike:

1. Organizacija podataka (kreiranje i održavanje baza bioloških informacija) i razvoj alata za lak pristup i upravljanje, kao i unošenje novih podataka.
2. Razvoj alata, algoritama, matematičkih formula i statističkih metoda, za analizu i detekciju relacija između podataka (poravnavanje sličnih proteina, generisanje filogenetskog stabla, predviđanje strukture i funkcije proteina i RNK sekvenci, razvoj proteinskih modela).
3. Interpretacija rezultata analiza na biološki razumljiv način.

Bioinformatika formuliše koncepte biologije vezane za molekule (u smislu fizičke hemije) primenjujući informatičke tehnike (izvedene iz disciplina kao što su: primenjena matematika, računarske nauke i statistika) za razumevanje informacija vezanih za ove molekule, i organizaciju informacija velikih razmera. Ukratko, bioinformatika se može tretirati kao informacioni sistem upravljanja u molekularnoj biologiji koji ima mnoge praktične primene [4].

Postoji više različitih tipova informacija koje bioinformatika obrađuje. Najčešća su tri primarna izvora podataka: (i) nukleotidne i proteinske sekvence, (ii) strukture makromolekula, (iii) rezultati eksperimenata među kojima su: ekspresija gena, metabolički putevi, regulatorne mreže, protein-protein interakcije. Pored toga, postoje i razni literaturni podaci. Na osnovu toga se vidi da postoji velika razlika u veličini i kompleksnosti baza podataka.

Zbog redundantnosti i mnogostrukosti podataka, kao i integracije među podacima, potrebna je organizacija informacija velikih razmera. Podaci se mogu grupisati na više načina prema biološkim sličnostima. Geni se mogu grupisati prema: sličnosti u sekvencama, sličnim biološkim funkcijama ili metaboličkim putevima, kao i na osnovu sličnosti u strukturi. Proteini međusobno mogu biti: (i) homologi (slične strukture i sekvence), (ii) analogi (slične strukture ali različite sekvence). Homologi mogu biti (i) ortologi (u različitim vrstama evoluirali su od istog potomka, zbog čega imaju istu funkciju), (ii) paralogi (nastali dupliranjem u okviru genoma i zato imaju različite, ali povezane funkcije). Kako postoje razni izvori i vrste informacija koje je potrebno kombinovati (zapis sekvenci, 3D koordinate, funkcije proteina, interakcije sa drugim proteinima, bibliografski podaci itd.), različite baze podataka se povezuju spoljnim vezama, a određeni internet servisi omogućuju integrisan pristup kroz više izvora podataka.

Bioinformatika je multidisciplinarna nauka koja integriše razvoj informatičkih i računarskih nauka primenjenih u biotehnološkim i biološkim naukama. Bioinformatika je trenutno u velikoj ekspanziji i razvoju, kao visoko interdisciplinarna nauka koja koristi tehnike i koncepte informatike, statistike, matematike, hemije, biologije, biohemije, fizike i lingvistike. Cilj bioinformatike je da izvede znanje iz bioloških podataka koristeći računarske analize, gde su podaci najčešće informacije zapisane u genetskom kodu, eksperimentalni rezultati iz raznih izvora, statistike pacijenata, i stručna literatura [5].

Postoje tri nivoa analize u bioinformatici:

1. Analiza pojedinačne nukleotidne ili proteinske sekvence (filogenetska analiza, evoluciona povezanost, predikcija subćelijske lokalizacije, povezanost sa drugim poznatim genima ili proteinima, predviđanje sekundarne i tercijalne strukture, itd.)

2. Analiza kompletnih genoma (identifikacija koja familija gena je prisutna a koja fali, lokacija gena na hromozomima, korelacija sa funkcijom ili evolucijom, duplikacija/ekspanzija familija gena, prisutnost ili odsutnost biohemijskih puteva, identifikacija nedostajućih enzima, događaji u evoluciji organizama, itd.)
3. Analiza gena i genoma u odnosu na podatke o funkcijama (analiza ekspresija gena i *mikroerej* (eng. *microarray*) podataka, proteomika, upoređivanje i analiza biohemijskih puteva, identifikacija gena koji učestvuju u specifičnim procesima, itd.)

1.1.2. Definicija bioinformatike

Računarska biologija je disciplina usko povezana sa bioinformatikom. Vrlo su slične i postoje značajna preklapanja, ali su to ipak dve odvojene discipline.

BISTIC komitet (*Biomedical Information Science and Technology Initiative Consortium*) Američkog nacionalnog instituta za zdravlje (*National Institutes of Health, NIH*), koji je specijalno oformljen za definiciju pojma bioinformatike i računarske biologije, dao je sledeću definiciju [6]:

> Bioinformatika i računarska biologija su ukorenjene u bionaukama (nauke o živim sistemima), baš kao i računarske i informatičke nauke i tehnologije. Oba ova interdisciplinarna pristupa su nastala iz specifičnih disciplina kao što su: matematika, fizika, računarstvo, biologija i bihejvioralne nauke. Bioinformatika i računarska biologija, svaka pojedinačno, imaju bliske interakcije sa bionaukama, kako bi razvile svoj pun potencijal. Bioinformatika koristi principe informatičkih nauka i tehnologija kako bi široke i kompleksne podatke iz bionauka učinila razumljivim i korisnim. Računarska biologija koristi matematičke i računarske pristupe da ukaže na određena teoretska i eksperimentalna pitanja u biologiji. Iako su bioinformatika i računarska biologija odvojene, postoje značajna preklapanja u okviru njihovih aktivnosti.
>
> Iako nijedna definicija ne može kompletno da eliminiše preklapanja u aktivnostima ili isključi varijacije u interpretacijama različitih individua ili organizacija, definicije bioinformatike i računarske biologije su:

Bioinformatika je naučna disciplina koja se bavi istraživanjem, razvojem ili primenom računarskih alata i pristupa za širu primenu bioloških, medicinskih, bihejvioristiških ili zdravstvenih podataka, a koja uključuje metode za prikupljanje, skladištenje, organizovanje, arhiviranje, analizu i vizualizaciju podataka.

Računarska biologija je naučna disciplina koja obuhvata razvoj i primenu teoretskih metoda i metoda za analizu podataka, tehnika za matematičko modelovanje i računarsku simulaciju, kao i za korišćenje u izučavanju bioloških, bihejvioralnih i socijalnih sistema.

1.1.3. Podoblasti bioinformatike

Osnovne grane bioinformatike su [7]:

- Genomika

  Genomika je poddisciplina u kojoj se vrši sekvenciranje, sastavljanje i analiza funkcija i struktura genoma (genom je kompletni skup DNK sekvenci u jednoj ćeliji nekog organizma). Osnovni zadaci genomike su: predviđanje gena, poravnavanje dve ili više sekvenci, identifikacija transkripcionih faktora, pregledanje genoma.
- Proteomika

  Proteomika se bavi analiziranjem funkcija i struktura proteoma (proteom je kompletan skup proteinskih sekvenci izraženih određenim genomom, ćelijom, tkivom ili organizmom). Osnovni zadaci proteomike su: poravnavanje dve ili više sekvenci, proučavanje funkcionalne konformacije proteina, određivanje aktivnog mesta, analiza protein-protein interakcije, subćelijska lokalizacija proteina i sortiranje proteina.
- Racionalni dizajn lekova (eng. *Rational drug design*)

  Razne računarske tehnike se koriste za pretraživanje molekula kao potencijalnih lekova i dizajniranje novih lekova. Time se skraćuje vreme i smanjuju troškovi procesa identifikacije bolesti i puštanja efektivnog leka na tržište.
- Baze bioloških podataka i analiza podataka (eng. *Bio Data Bases and Data Mining*)

Uz ogromne količine podataka u bazama nukleotidnih i proteinskih sekvenci i 3D strukturama proteina, javlja se potreba za razvojem novih metoda za analizu tih podataka, u kojima se koriste principi metoda iz oblasti analize podataka (eng. *data mining*).

- Molekularna filogeneza

  Molekularna filogeneza određuje kvantitativan kriterijum za klasifikaciju organizama preko molekularne, odnosno bioinformatičke analize proteinskih i nukleotidnih sekvenci.

- Mikroerej Informatika

  Mikroerej je skup čipova koji se koriste za analizu ekspresije gena. Kako nisu svi geni u ćeliji aktivni sve vreme, ekspresija gena određuje koji geni i kada su aktivni u ćeliji. Mikroerej čipove čini matrica uzoraka DNK fragmenata uzetih u različitim vremenima (stanjima ćelije) koji su ofarbani različitim bojama prema vremenima. Rezultat je slika određene rezolucije, koju je potrebno računarskim metodama obraditi, preprocesirati, analizirati metodama klasifikacije i izvesti statistički validne zaključke.

- Sistemska biologija

  Sistemska biologija se bavi modelovanjem bioloških procesa na ćelijskom nivou. Identifikacijom osnovnih komponenti, parametara i promenljivih, kroz modelovanje jednačina, ona objašnjava osnovna pitanja u biologiji. Ukratko, sistemska biologija predstavlja *in silico* rekonstrukciju bioloških fenomena.

1.1.4. Bioinformatičke metode

Osnovni zadaci bioinformatike (u istraživanju) su [7, 8]:

- Poravnavanje sekvenci (eng. S*equence alignment*) - uređenje odnosno poravnavanje dve ili više nukleotidnih ili proteinskih sekvenci za identifikaciju sličnih oblasti po homologiji, a koje mogu biti posledica funkcijske, strukturne ili evolucijske povezanosti sekvenci;
- Predviđanje gena (eng. G*ene finding*) - identifikacija regiona DNK koji kodiraju gene;

- Sastavljanje genoma (eng. *Genome assembly*) - ponovno sastavljanje velikog broja kratkih DNK sekvenci u originalan hromozom iz koga potiču;
- Računarski potpomognut dizajn lekova (racionalni dizajn lekova) - inventivan proces otkrivanja novih medicinskih lekova na osnovu poznavanja biološke mete;
- Otkrivanje lekova - proces kojim se otkrivaju ili dizajniraju lekovi, a koji obuhvata: identifikaciju kandidata, visokopropusni skrining, razvoj leka, sintezu, karakterizaciju, klinička ispitivanja i testiranje terapeutske efikasnosti;
- Strukturno poravnavanje proteina - uspostavljanje homologije između dve ili više polimerske strukture, na osnovu njihovih oblika i trodimenzionalne konformacije i najčešće se koristi za proteinske tercijalne strukture, pogotovo kada proteini poseduju malu sličnost sekvenci pa se korišćenjem metoda poravnavanja sekvenci ne može detektovati njihov evolutivni ili funkcionalni odnos;
- Predviđanje strukture proteina (proteinsko savijanje) - predviđanje trodimenzionalne, odnosno sekundarne, tercijalne i kvaternarne strukture na osnovu primarne strukture proteina, predstavlja jedan od bitnijih ciljeva u medicini (dizajnu leka) i biotehnologiji (dizajnu novih enzima);
- Predviđanje ekspresije gena (ekspresija gena je proces u kome se na osnovu informacije iz gena sintetiše funkcionalni genski produkt tj. protein ili RNK);
- Protein-protein interakcije - analiza interakcije između proteina (interakcija među proteinima se javlja kada se proteini vežu da bi vršili svoju biološku funkciju);
- Identifikacija mesta vezivanja transkripcionog faktora (eng. *Transcription factor binding site identification*) - identifikacija mesta na određenoj DNK sekvenci gde se odvija vezivanje sa transkripcionim faktorom (proteinom);
- Subćelijska lokalizacija proteina (eng. *Protein sub-cellular localization*) i sortiranje proteina (eng. *Protein sorting*) - određivanje dela ćelije u kojem određen protein egzistira i lociranje pozicije gde se transportuje protein u ćeliji;
- Izučavanje genomskih asocijacija (eng. *Genome-wide association study, GWAS*) - analiza zajedničkih genetskih varijanti kod više osoba za određivanje varijante vezane za datu osobinu, a specifično se koristi za vezu između bolesti i polimorfizama pojedinačnih nukleotida (eng. *Single-nucleotide polymorphism, SNP*);

- Pregledanje genoma (eng. *Genome browsing*) - razvoj metoda za kompresovanje i jasnu vizualizaciju DNK sekvenci;
- Određivanja aktivnog mesta (eng. *Active site determination*) - određivanje regiona proteina koji je hemijski najaktivniji;
- Modelovanje evolucije - generisanje modela evolucije (evolucija je promena nasleđenih karakteristika populacija kroz sukcesivne generacije).

1.2. Hemoinformatika

1.2.1. Opšte o hemoinformatici

Hemoinformatika je naučna disciplina koja primenjuje informatičke metode za rešavanje hemijskih problema. Veoma je mlada disciplina, nastala 1960-tih godina. Za prvo spominjanje pojma hemoinformatike zaslužan je Frenk Braun [9]:

Korišćenje informacionih tehnologija postalo je bitan deo u procesu otkrivanja lekova. Hemoinformatika se sastoji iz transformacije podataka u informacije, a informacija u znanje, u svrsi boljeg i bržeg odlučivanja u procesu identifikacije 'lead'[1] *jedinjenja za lek.*

Podaci u hemoinformatici su bilo koji rezultati dobijeni zapažanjima: fizička merenja, (ne)prisutnost reakcija, određivanje bioloških aktivnosti. Informacija nastaje kada se podatak stavi u kontekst sa drugim podacima. Mera biološke aktivnosti jedinjenja dobija vrednost informacije kada je poznata i struktura jedinjenja. Izvođenje znanja zahteva određen nivo apstrakcije. Zakonitosti modela se izvode iz niza posmatranja i analogno se izvode predviđanja. U slučaju hemoinformatike procesom apstrakcije se dobija znanje o osobinama jedinjenja. Fizički, hemijski i biološki podaci se povezuju međusobno i/ili sa podacima o strukturi jedinjenja, čime se dobijaju informacije. Te informacije se dalje analiziraju procesima apstrakcije, odnosno metodama induktivnog učenja na osnovu kojih se generiše model koji omogućava izvođenje predviđanja. Induktivno učenje je cikličan proces u kojem se iz niza

[1] *Lead* jedinjenje je hemijsko jedinjenje koje poseduje odgovarajuću biološku aktivnost, čija struktura predstavlja polaznu tačku za dalju modifikaciju u procesu otkrivanja lekova. Pogledati odeljak 1.3.1.

posmatranja (rezultata eksperimenata) izvodi skup zaključaka (model) na osnovu kojih se predviđaju nova zapažanja.

Za razliku od definicije Frenk Brauna sa naglaskom na dizajn lekova, mnogo širu definiciju hemoinformatike dao je Greg Pariz [10]:

Hemoinformatika *je opšti pojam koji objedinjuje dizajn, kreiranje, organizaciju, pristup, upravljanje, analizu, vizualizaciju i korišćenje hemijskih informacija.*

Hemija generiše ogromnu količinu podataka i taj broj raste velikom brzinom. Poznato je preko 100 miliona hemijskih jedinjenja [3] i preko 300 hiljada 3D struktura organskih jedinjenja dobijenih metodama rentgenske strukturne analize i oko 200 hiljada infracrvenih spektara. Takođe, mnogi problemi u hemiji su suviše kompleksni da bi se rešili običnim matematičkim metodama. Zato je za skladištenje, upravljanje i obradu hemijskih podataka neophodno primeniti informatičke i računarske metode.

1.2.2. Hemoinformatika i bioinformatika

Bioinformatika se uglavnom bavi analizom gena i proteina, koji su takođe hemijska jedinjenja, pa time i predmet analize hemoinformatike. Zato ne postoji jasna razlika između hemoinformatike i bioinformatike. Dok se hemoinformatika bavi hemijskim jedinjenjima, najčešće malim molekulima ali i velikim, bioinformatika se bavi isključivo makromolekulima, genima i proteinima, koji su takođe predmet analize i hemoinformatike. Zbog boljeg razumevanja struktura, osobina i funkcija proteina i nukleotidnih sekvenci, bioinformatika i hemoinformatika tesno sarađuju, pogotovo u dizajnu lekova.

Metode u genomici identifikuju proteinske mete za nove kandidate za lek. Sa druge strane pošto su lekovi najčešće mali molekuli, hemoinformatičke metode pronalaze nove *lead* strukture i optimizuju ih za kandidate za lek (slika 1.1). Iako su hemoinformatičke metode fokusirane na analizu malih i srednjih molekule, većina tih metoda se može primeniti i na makromolekule, odnosno proteine i nukleotidne kiseline.

Slika 1.1. Saradnja između bioinformatike i hemoinformatike.

1.2.3. Hemoinformatičke metode

Osnovni predmeti hemoinformatike, odnosno specijalizovane podoblasti (sa naglaskom na probleme i rešenja) su [11]:

- Predstavljanje hemijskih jedinjenja zapisom

  Razvijen je velik broj metoda za prezentaciju zapisa hemijskih jedinjenja na računaru. Posebno su razvijene metode koje jedinstveno opisuju hemijske strukture i njihove osobine kao što su prstenovi, aromatičnost, 3D struktura, molekulska površina.

- Predstavljanje hemijskih reakcija

  Pored početnih jedinjenja i rezultata u reakcijama, bitno je pravilno predstaviti i uslove reakcije, mesto reakcije i veze.

- Podaci u hemiji

  Hemija proizvodi širok raspon podataka, od fizičkih, hemijskih i bioloških, do binarnih podatak za klasifikaciju, realnih podataka za modelovanje i spektralnih sa visokom gustinom informacija. Sve te podatke je potrebno obraditi i dostaviti na pristupačan način za dalju razmenu i analizu.

- Izvori podataka i baze podataka

  Velika količina podataka u hemiji (jedinjenja, 3D strukture, reakcije, spektri, literatura itd.) dovela je do razvoja baza podataka za skladištenje ovih informacija u električnoj formi.

- Metode za pretraživanje struktura

  Razvijene su razne metode za pretraživanje baza podataka po celoj strukturi, podstrukturi, kao i po sličnosti u strukturama.

- Metode za izračunavanje fizičkih i hemijskih podataka

Ovim metodama se direktno izračunavaju raznovrsne fizičke i hemijske karakteristike jedinjenja, a posebno se kvantnom mehanikom mogu izračunati i proceniti razne osobine do određene preciznosti.

- Kalkulacija strukturnih deskriptora

  Mnoge karakteristike jedinjenja se ne mogu direktno izračunati, zbog čega se koristi indirektan pristup kojim se prvo struktura jedinjenja predstavlja strukturnim deskriptorima. Zatim se određuju zavisnosti između deskriptora i osobina analiziranih jedinjenja, na osnovu kojih se uz pomoć metoda induktivnog učenja predviđaju osobine početnog jedinjenja.

- Metode za analizu podataka

  Hemoinformatika koriste razne metode za analizu podataka: statističke metode, prepoznavanje šablona, veštačke neuralne mreže, genetski algoritmi itd.

- Model kvantitativnih odnosa strukture i dejstva/osobina aktivnih supstanci (eng. *Quantitative structure-activity/property relationship – QSAR/QSPR*)

  Veliki broj deskriptora dobijenih iz molekulskih struktura u poslednjih 40 godina, doveo je do razvoja modela kvantitativnih zavisnosti između hemijske strukture i bioloških aktivnosti supstanci (odnosno fizičko-hemijskih osobina) - QSAR (odnosno QSPR). Na osnovu ovog modela, koji može biti regresioni ili klasifikacioni model, predviđa se aktivnost novog jedinjenja.

- Hemometrija

  Hemometrija je naučna disciplina koja: (i) povezuje mere iz hemijskih sistema i procesa vezanih za stanja sistema, (ii) generiše optimalne procedure i eksperimente za merenje, (iii) izvlači maksimalnu informaciju iz analize hemijskih podataka, kroz primenu matematičkih i statističkih metoda. Koristi razne metode za analizu podataka: prepoznavanje šablona, veštačke neuralne mreže, obrada signala itd.

- Modelovanje molekula

  Modelovanje molekula se sastoji od: (i) vizualizacije 3D molekulskih modela i detaljne vizualizacije 3D strukture proteina, (ii) generisanje 3D struktura i modelovanje proteina, (iii) simulacije dinamike molekula.

- Računarom potpomognuto tumačenje struktura (eng. *Computer assisted structure elucidation, CASE*)

Primenom računarskih metoda i programa, generišu se sve moguće molekulske strukture koje su određene skupom podataka iz spektroskopskih metoda.

- Računarom potpomognut sintetički dizajn (eng. *Computer assisted synthesis design, CASD*)

  Uz pomoć računarskih metoda i tehnika, i na osnovu znanja o hemijskim reakcijama i reaktivnosti, vrši se sinteza organskih jedinjenja (sklapanje blokova molekula).

1.3. Dizajn lekova

Dizajn lekova je proces nalaženja novih lekova zasnovan na poznavanju biološke mete. Lek je najčešće ligand[2], koji u interakciji sa proteinom aktivira ili inhibira određenu funkciju proteina (biološke mete), čime se postiže terapeutski efekat. **Biološka meta**, **ciljno mesto** ili **target**[3] je biopolimer, protein ili nukleotidna kiselina, koji u interakciji sa aktivnim molekulom (lekom) menja svoju aktivnost i time utiče na sprečavanje bolesti. Aktivni molekul može vezivanjem za target: (i) promeniti konformaciju targeta čime menja njenu funkciju (agonisti), (ii) sprečiti druge supstance (hormone) da se vežu za target, bez promene konformacije ili funkcije targeta (antagonisti, inhibitori, blokatori). Proteinski targeti najčešće mogu biti: enzimi, G protein-spregnuti receptori (GPCR), jonski kanali, membranski transporteri, nuklearni hormonski receptori, strukturni proteini [12]. Dizajn lekova se najčešće oslanja na računarsko modelovanje.

1.3.1. Proces otkrivanja lekova

Otkrivanje lekova je jedna od osnovnih oblasti primene hemoinformatike. Razvoj novog leka je vremenski dugotrajan i skup proces, koji traje od 10 do 15 godina i košta preko milijardu dolara da bi se lek pustio na tržište [13, 14]. Zbog toga je bitan

[2] Ligand je mali molekul koji se čvrsto vezuje za svoj ciljni protein na biomolekulu s kojim formira kompleks. Pogledati odeljak 2.1.1.

[3] Target je najčešće korišćen termin u stručnoj literaturi na srpskom jeziku.

razvoj *in silico* metoda koji ubrzavaju i pojeftinjuju proces otkrivanja lekova, kao alternativa za *in vitro* testove, čime se u ranoj fazi smanjuje broj jedinjenja sa nepovoljnim farmakološkim osobinama. Loše ADMET osobine leka su razlog za osipanje kandidata u toku procesa razvoja leka.

Glavni razlozi za osipanje novih lekova su [13]:

- Nedostatak kliničke efikasnosti
- Neodgovarajuća farmakokinetika (Farmakokinetika ja nauka koja prati sudbinu leka u organizmu, tj. procese resorpcije, distribucije, metabolizma i ekskrecije lekova)
- Toksičnost na životinjama
- Komercijalni razlozi
- Problemi formulisanja

Procenjeni procenat osipanja lekova je i do 90% [15], što utiče na diskontinuitet u procesu razvoja i stvara finansijske gubitke. Zbog toga je veoma bitno detektovati ADMET probleme u što ranijoj fazi primenom hemoinformatičkih metoda za predviđanja farmakoloških osobina, zajedno sa identifikacijom *lead* strukture.

Proces otkrivanja lekova se sastoji od sledećih koraka [13]:

1. Identifikacija targeta

   Analizom procesa bolesti, proces se deli na bitne komponente, i zatim se identifikuje ona koja je najbitnija za manifestaciju bolesti. Cilj identifikacije ciljnog mesta je razumevanje bioloških procesa vezanih za bolest i identifikacija mehanizama i struktura pojedinačnih elemenata bolesti. Najčešći targeti su: receptori (45%) i enzimi (28), ostatak čine hormoni (11%), jonski kanali (5%), DNK (2%), receptori jezgra (2%), nepoznato (9%) [16]. Na osnovu znanja o ljudskom genomu, postoji preko 30 hiljada potencijalnih meta od kojih je poznato samo oko 3 hiljade koji su povoljni za lek [17, 18].

2. Validacija targeta

   Osobine ciljnog mesta se analiziraju *in vitro* (na ćelijskim kulturama) i u *in vivo* modelima (na modelima bolesti kod životinjama). Target je uspešno prošao proveru kada određeno delovanje na metu pokaže povoljne efekte u modelu

bolesti. Takođe je bitno uzeti u obzir da modifikacija targeta ubije mikroorganizam (inhibira virulentnost), a nema uticaja na domaćina.

3. Identifikacija *lead* jedinjenja

   Lead jedinjenje je hemijsko jedinjenje koje poseduje odgovarajuću biološku aktivnost, čija struktura predstavlja polaznu tačku za dalju modifikaciju u cilju optimizacije aktivnosti, selektivnosti, fizičko-hemijskih osobina i drugih parametara. Na početku procesa pronalaženja *lead* jedinjenja, dizajniraju se i sintetišu biblioteke jedinjenja. Zatim se vrši skrining (pregledanje, sortiranje) biblioteka kroz eksperimentalne testove, korišćenjem identifikovane mete. Ona jedinjenja koje pokažu aktivnost u testovima se označavaju kao pogoci (eng. *hits*), i nastavlja se njihova provera. Lek mora biti siguran, efikasan i da poseduje: mogućnost dobre apsorpcije u organizmu, metabolizam koji mu omogućava dovoljno dug poluživot, selektivan efekat na metu, malu toksičnost i minimalne neželjene efekte.

4. Optimizacija *lead* jedinjenja

   Optimizacija je iterativan proces koji sistematski modifikuje i razvija jedinjenje koje ima visoku potentnost, visoku selektivnost, određen farmakokinetički profil, i minimalnu toksičnost i mutagenost.

5. Pretklinička istraživanja

   Pretklinička i klinička ispitivanja traju u proseku 10 godina. Pretklinička ispitivanja se obavljaju *in vitro* i *in vivo* na životinjama, čime se procenjuju biološke aktivnosti novog jedinjenja.

6. Klinička istraživanja

   U prvoj fazi kliničkog ispitivanja ispituje se bezbednost i određuje potrebna doza na uzorku od 20 do 100 dobrovoljaca. Druga faza je fokusirana na bezbednost, efikasnost i sporedne efekte koji se ispituju na grupi od 100 do 300 dobrovoljaca. U trećoj fazi, u kojoj učestvuje oko hiljadu dobrovoljaca, ispituje se i dokazuje efikasnost i bezbednost leka na duži vremenski period korišćenja.

7. Odobrenje za lek

   Pre puštanja leka na tržište, potrebno je da lek prođe proces provere od strane nadležne agencije. Agencija za evaluaciju medicinskih proizvoda u Evropi je

EMA (*European Medicines Agency*), dok je u Americi FDA (*Food and Drug Administration*).

1.3.2. Dizajn lekova primenom hemoinformatičkih metoda

Polja primene hemoinformatike u procesu dizajna lekova su [13]:

- Selekcija podskupa i pretraživanje po sličnosti/različitosti

  Biblioteke hemijskih jedinjenja sadrže oko 10^6 jedinjenja od kojih se 10^3 pusti na tržište u obliku komercijalnih lekova, što čini samo oko 0.1% baze svih jedinjenja. Zbog toga je bitan razvoj moćnih računarskih tehnika za selekciju podskupa, kako za HTS tako i za virtuelni skrining. Metode koje analiziraju raznovrsnost selektovanih podskupova obezbeđuju pokrivenost hemijskog prostora jedinjenja, u kojima su rastojanja/sličnosti među jedinjenjima definisana 2D i 3D deskriptorima, kao i osobinama površina molekula.

- Analiza podataka visokopropusnim skriningom (eng. *high-throughput screening, HTS*)

  Visokopropusni skrining je eksperimentalna metoda za brzo merenje efekata i aktivnosti supstanci u biološkim ili hemijskim eksperimentima, koja koristi automatizovane procese i načine za praćenje više testova istovremeno. Sa sve većom količinom prikupljenih jedinjenja sintezom iz kombinujućih biblioteka, a zatim metodom HTS, javlja se potreba za razvojem matematičkih i računarskih alata za njihovo pretraživanje i izvođenje bitnih informacija. Za identifikaciju povezanosti i šablona u hemijskih bazama se najčešće koriste metode analize podataka (eng. *data mining*).

- Virtuelni skrining

  Virtuelni skrining (VS) ili *in silico* skrining je metoda selekcije jedinjenja izračunavanjem njihovih bitnih osobina u računarskim modelima [19]. Te osobine su: visoka potentnost, selektivnost, određene farmakokinetičke osobine i minimalna toksičnost. Virtuelni skrining je brži i jeftiniji od HTS jer nema troškova za eksperimentalne sinteze i testove, pa se koristi pre ili paralelno sa HTS. U procesu VS dizajna leka mogu se primeniti metode zasnovane na ligandu i na strukturi (slika 1.2).

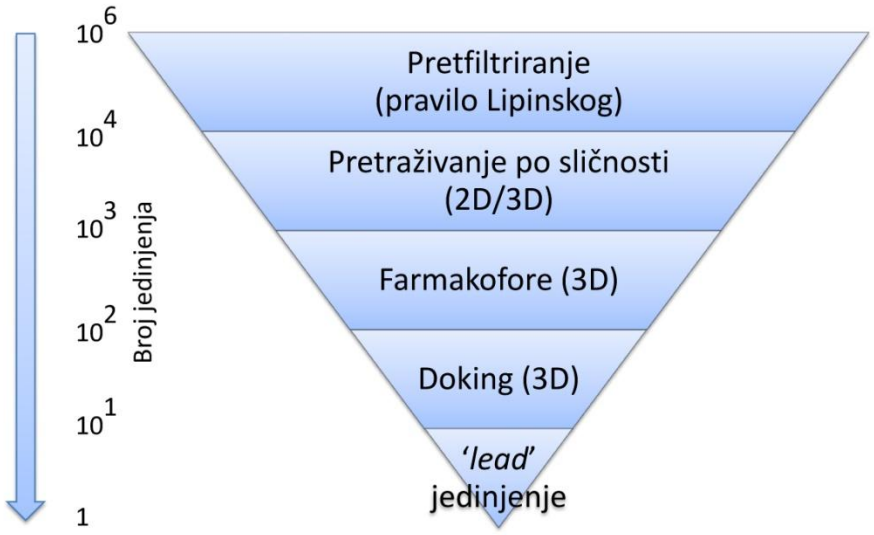
- Dizajn kombinujućih biblioteka

  Podaci dobijeni iz HTS i VS metoda se mogu primeniti za generisanje kombinujućih biblioteka jedinjenja. Najčešće korišćena metoda za kreiranje kombinovanih jedinjenja je genetski algoritam.

- Ostali zadaci

  Pored osnovnih primena, hemoinformatika se primenjuje i u sledećim problemima dizajna lekova:

  - Razvoj QSAR/QPSR modela
  - Fleksibilno 3D poravnavanje
  - Predviđanje raznih fizičko-hemijskih osobina (rastvorljivost, lipofilnost, broj rotirajućih veza, polarna površina, farmakokinetičke osobine, ADMET profil, povoljnost za lek, povoljnost za *lead* jedinjenje itd.)



Slika 1.2. Tok procesa virtuelnog skrininga, od pripreme podataka do otkrivanja novih *lead* jedinjenja.

1.4. Filogenetska analiza

Evolucija je promena naslednih karakteristika bioloških populacija kroz uzastopne generacije. Nasleđene osobine su ekspresije gena i menjaju se usled mutacija, rekombinacija ili protoka gena. Evolucija se pojavljuje kada se prirodnom selekcijom ili genetskim driftom promeni učestalost starih, novim osobinama.

Filogenija (filogeneza) definiše evolutivni odnos u okviru familije blisko povezanih sekvenci. Ti odnosi se prikazuju evolutivnim stablom. Svaki čvor se smatra taksonomskom zajednicom, pri čemu su listovi trenutno postojeći, a unutrašnji čvorovi hipotetički organizmi. Dužine grana predstavljaju procenjeno vreme kroz koje su se grupe organizama razvijale.

Molekularna filogeneza je grana filogenije koja dolazi do informacija o evolutivnim odnosima organizama analizom naslednih molekulskih razlika u nukleotidnim sekvencama i proteinima.

Metode za konstruisanje i analizu filogenetskih stabala se mogu klasifikovati u [20, 21]:

- Metode zasnovane na rastojanjima

  U prvom koraku metoda zasnovanim na rastojanjima, izračunaju se rastojanja između svake dve sekvence, gde je rastojanje definisano u zavisnosti od mere i evolutivnog modela [22-25]. U drugom koraku matrica rastojanja se transformiše u stablo koristeći neki od sledećih algoritama za klasterisanje:

  - UPGMA metoda (eng. *Unweighted Pair Group Method with Arithmetic Mean, UPGMA*) [26, 27]
  - Metoda spajanja suseda (eng. *Neighbor-Joining, NJ*) [28]
  - Fič-Margoliš metoda (eng. *Fitch-Margoliash*) [29]

- Metode zasnovane na karakterima

  Metode zasnovane na karakterima pronalaze najbolje stablo izračunavanjem pogodnosti različitih topologija, gde se izračunavanja vrše na svakom pojedinačnom aminokiselinskom ostatku sekvence (karakteru), odnosno na svakoj poziciji. U odnosu na kriterijum optimalnosti, metode se mogu podeliti u:

  - Metode maksimalne parsimonije (eng. *Maximum Parsimony, MP*) [30-32]
  - Metode zasnovane na verovatnoći:

- Metoda maksimalne verodostojnosti (eng. *Maximum Likelihood, ML*) [33], sa strategijom da se minimizuje broj evolutivnih promena tj. supstitucija
- Bajesovo izvođenje (eng. *Bayesian inference*) [34], koje je zasnovano je na Monte-Karlo uzorkovanju [35]

U zavisnosti od nivoa sličnosti sekvenci, MP metode se najčešće koriste za veoma slične sekvence, metode zasnovane na rastojanjima se primenjuju na grupi sekvenci koje dele prepoznatljivu sličnost, a probabilističke metode za grupu koja poseduje nizak nivo sličnosti [36].

Metode zasnovane na rastojanju i njihove implementacije (Fitch, Kitsch, Neigbor tools, u paketu Phylip [37], ClustalW i ClustalX [38], MEGA [39]) kao i MP metode (Prot-pars [37], PAUP [40]) su veoma brze i pomoću njih se mogu analizirati veoma velike grupe sekvenci. Za razliku od njih, metode zasnovane na verovatnoći (ProtML [41], Tree-PUZZLE [42], PAML [43], MrBayes [44], BEAST [45], PHYML [46]) daju tačnije rezultate, ali su računarski veoma zahtevne [47-49].

Svi ovi pristupi zahtevaju višestruko poravnavanje sekvenci (eng. *Multiple Sequence Alignment, MSA*) i u većini se pretpostavlja određeni evolutivni model i empirijske matrice supstitucije. Kvalitet izvedenog filogenetskog stabla zavisi od kvaliteta MSA [50]. Problemi vezani za MSA su: vremenska kompleksnost, ograničen broj sekvenci, izbor parametara za MSA programe [49, 51], kontraverzni evolutivni modeli, dvosmislenost kriterijuma za ocenu MSA, različiti alati generišu različita poravnavanja [52]. U popularnim progresivnim MSA programima (ClustalW [53], T-Coffee [54], MAFFT [55]) pouzdanost MSA je vezana za poravnavanje prve dve najsličnije sekvence [54]. Da bi se prevazišli ovi problemi, poslednjih godina razvijeno je nekoliko metoda za filogenetsku analizu proteinskih sekvenci, nezavisnih od MSA, koje su zasnovane na različitim rastojanjima između sekvenci: *Bhattacharyya* rastojanje [56], Lempel-Ziv kompleksnost [57, 58], vektori odlika (*feature vectors*) [59], mera relativne kompleksnosti [51, 60], metoda za veoma divergentne sekvence PHYRN [49].

2. Međumolekulske interakcije u biološkom sistemu

Prvi i najbitniji korak u procesu otkrivanja lekova je detekcija bioloških meta (targeta), koji se sastoji od identifikacije i ranog testiranja terapeutskih targeta. Taj korak traje u proseku od 12 do 15 godina i košta oko 1.8 milijardi dolara [14]. Zato je veoma bitno razumeti proces međusobnog prepoznavanja biomolekula, odnosno međumolekulskih interakcija, koji je osnova identifikacije proteinskih i nukleotidnih targeta i predikcije interakcija lek-protein i lek-nukleotid, što bi omogućilo skraćenje perioda i smanjenje troškova procesa.

2.1. Kratkodosežne međumolekulske interakcije

Osnova današnjeg shvatanja međumolekulskih interakcija u biološkim sistemima zasniva se na hipotezi „ključ-brava“. Ova hipoteza o komplementarnim površinama između interreagujućih molekula, koju je 1894. godine predložio Emil Fišer (*Hermann Emil Fischer*) dobitnik Nobelove nagrade za hemiju 1904. godine, zajedno se teorijom sudara, polaze od dve osnovne pretpostavke [61]:

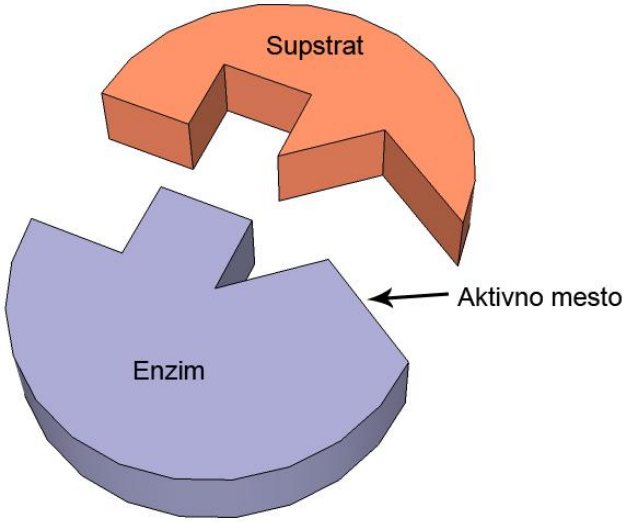
(i) Prvi kontakt između interreagujućih molekula se dešava slučajno kao posledica njihovog termalnog kretanja.
(ii) Za njihovu dalju interakciju odgovorne su slabe kovalentne veze koje deluju na rastojanjima manjim od 2Å.

Fišerova hipoteza i njene kasnije modifikacije poslužile su kao osnova za razvoj brojnih bioinformatičkih alata za analizu veza između strukture i funkcije bioloških molekula i njihovih međusobnih interakcija.

2.1.1. Koncept „Ključ-brava“

Model ključ-brava je model interakcije između enzima i supstrata, koji tvrdi da enzim i supstrat poseduju **komplementarne** geometrijske oblike koji se uklapaju tačno

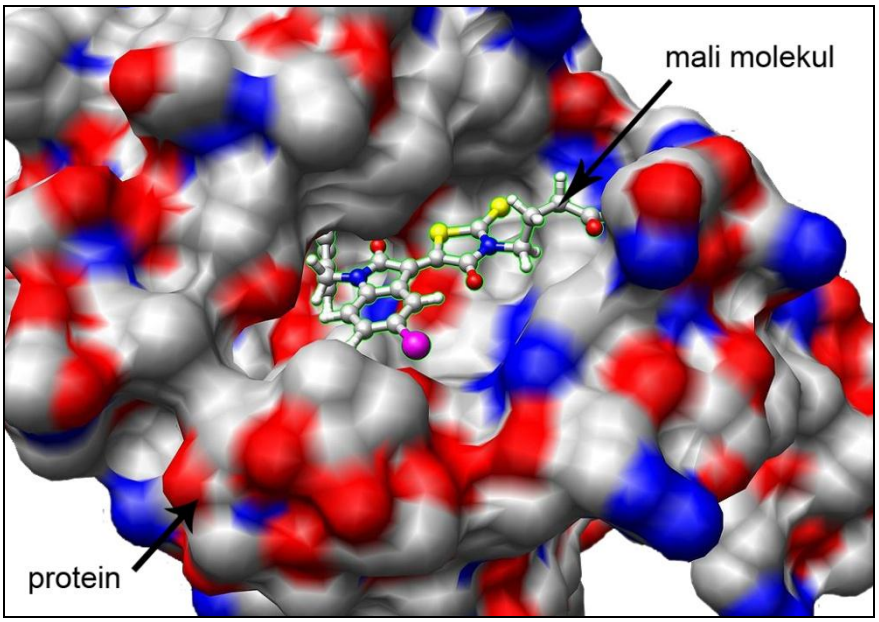
jedan u drugi (slika 2.1). Baš kao što ključ ulazi u bravu, samo supstrat (ključ) određenog oblika i veličine može se uklopiti u aktivno mesto (rupu za ključ) enzima (brave). Aktivno mesto je specifičan region enzima na koji se vezuje supstrat i ima jedinstveni geometrijski oblik koji je komplementaran sa oblikom molekulskog supstrata.



Slika 2.1. Princip ključ-brava. Supstrat i enzim komplementarnih oblika interreaguju i formiraju stabilan kompleks.

Model ključ-brava se posebno koristi u **molekulskom dokingu** (eng. *Molecular docking*). Doking je proces u kome se dva molekula vezuju u stabilan kompleks. Metode za doking predviđaju 3-dimenzionalne orijentacije molekula u doking procesu. Informacija o prioritetnoj orijentaciji se koristi za predviđanje snage i afiniteta vezivanja dva molekula. Orijentacije dva interreagujuća molekula utiču na tip proizvedenog signala, odnosno agonizam ili antagonizam.

Ako se protein posmatra kao brava, a ligand kao ključ (**ligand** je mali molekul koji se čvrsto vezuje za svoj ciljni protein na biomolekulu s kojim formira kompleks), doking se može definisati kao problem optimizacije, koji za zadatak ima pronalaženje najbolje konformacije i orijentacije liganda koji se vezuje na određeni protein (slika 2.2). Računarski alati za doking su zasnovani na simulaciji procesa molekulskog prepoznavanja (i) optimizacijom konformacije proteina i liganda u komplementarne strukture, ili (ii) minimizacijom ukupne energije u protein-ligand interakciji.



Slika 2.2. Mali molekul vezan za protein. Izvor: Wikipedia-Docking.

2.1.2. Računarske metode za određivanje interakcije na bazi 3D struktura

2.1.2.1. Proteinska banka podataka

Proteinska banka podataka (*Protein Data Bank - PDB*) je kolekcija trodimenzionalnih struktura velikih bioloških molekula (proteini i nukleotidne kiseline). PDB je formirao Volter Hamilton (*Walter Hamilton*) 1971. godine u Brookhaven nacionalnoj laboratoriji u Teksasu, a 2003. godine je transformisana u internacionalnu organizaciju **wwPDB** [62], koju sačinjavaju 4 članice zadužene za prikupljanje, obradu i distribuciju PDB podataka: RCSB PDB iz Amerike, PDBe iz Evrope, PDBj iz Japana i BMRB iz Amerike. Cilj wwPDB je održavanje jedinstvene slobodno dostupne PDB arhive podataka o strukturi makromolekula.

Podaci o 3D strukturi molekula se dobijaju primenom: (i) rendgenske strukturne analize (eng. *X-ray diffraction, XRD*), (ii) NMR spektroskopije (eng. *Nuclear Magnetic Resonance, NMR*), (iii) elektronske mikroskopije, (iv) hibridnih metoda i (v) ostalih metoda. U julu 2013. godine PDB baza je sadržala preko 92 hiljade struktura [63].

2.1.2.2. Metode za molekularni doking

Tri osnovne komponente u metodama za doking su:

1. Reprezentacija sistema, koga čini definisanje matematičkog modela koji opisuje površinu molekula. U zavisnosti od konformacije liganda, model može biti čvrst [64] ili fleksibilan [65, 66]
2. Pretraživanje prostora konformacija, u kojem su kritični elementi brzina i efektivno pokrivanje prostora konformacija
3. Rangiranje potencijalnih rešenja, gde je bitno je definisati dobru skor-funkciju koja ima ulogu izračunavanja i rangiranja velikog broja potencijalnih rešenja i razdvajanje prirodnih od neprirodnih konformacija. Skor-funkcija je brza aproksimativna matematička funkcija za predviđanje interakcija između dva molekula kada se vežu i formiraju stabilan kompleks.

Najčešći modeli za reprezentaciju površina koriste:

- Opis geometrijskih osobina [67]. U ovu grupu spadaju model rasejanih kritičnih tačaka [68, 69] i modelovanje sferama [70].
- Zapreminske osobine [71, 72], kod kojih se koristi metoda kreiranja površinske mreže tačaka.

Algoritmi za pretraživanje se mogu podeliti na [67, 73]:

- Algoritme stohastičke optimizacije

  Uzorkovanje na prostoru konformacija molekula se vrši pojedinačnom promenom na ligandu. U ovu grupu spadaju:
  - Monte Karlo metoda
  - Genetski algoritmi (evolutivno programiranje)
  - Tabu pretraživanje
  - Simulirano kaljenje
- Sistematsko pretraživanje

  Pretraživanje se vrši po svim stepenima slobode u molekulu. Ova grupa uključuje algoritme:
  - Pretraživanje fragmenata

Postepeno se pravi ligand na aktivnom mestu, vezujući više fragmenata.

- Pretraživanje konformacija

  Generišu se sve moguće kombinacije konformacija rotacijom svih rotirajućih veze i uzimajući se u obzir tačke komplementarnosti između molekula.

- Geometrijska rastojanja

  Aktivna mesta su prikazana sferom i uzimaju se u obzir samo vodonične veze.

- Algoritmi zasnovani na bazi podataka

  Ove metode kombinuju pregenerisane konformacije iz baza struktura jedinjenja.

- Simulacije
  - Molekularna dinamika

    Izračunavaju se rešenja sistema Njutnovih jednačina kretanja i traži se minimum energije kompleksa vezanih molekula.

  - Minimizacija energije

    Nalazi se lokalni minimum energije. Ove metode se koriste kao komplementarne drugim metodama pretraživanja.
- Kombinovane (hibridne) metode

Postoje tri osnovne klase skor-funkcija procene i rangiranja [67, 74, 75] zasnovane na:

1. Polju sila (eng. *force field*), koje koriste elektrostatičke i Van der Valsove sile za procenu interakcije (AMBER [76], OPLS [77], CHARMM [78], ECEPP/3 [79], GROMOS [80], G-Score [81], D-Score [82])
2. Empirijskim zapažanjima, koje koriste eksperimentalno utvrđene energije vezivanja (ChemScore [83], LUDI [84], F-Score [85], Fresno [81])
3. Znanju, gde su skor-funkcije statistički zasnovane na eksperimentalno utvrđenim 3D strukturama molekula (DrugScore [86], SMoG [87], PMF [88]).

Pored ove tri osnovne klase skor-funkcija, postoje i konsenzus funkcije koje kombinuju informacije iz različitih (gore navedenih) osnovnih klasa funkcija (X-SCORE [89]).

2.1.2.3. Programi za molekularni doking

U metodama za doking se koriste **veoma zahtevni** računarski algoritmi (genetski algoritam, iscrpno pretraživanje, Monte Karlo simulacija, pseudo-Braunovo uzorkovanje, simulirano kaljenje, metoda najstrmijeg opadanja itd.), zbog velikog broja kombinacija za spajanje dva molekula. Postoje tri translaciona i tri rotaciona stepena slobode, a broj mogućnosti eksponencijalno raste sa veličinom komponenti [90]. Predviđanje afiniteta vezivanja (snage interakcije) liganda za receptor je vrlo komplikovan zadatak. Pri izračunavanju afiniteta vezivanja potrebno je analizirati razne termodinamičke osobine liganda (entapliju, entropiju, elektrostatički potencijal molekula, dipolni momenat, efekat odbijanja, pridodate molekule vode, protonaciju, itd.), faktore koji utiču na termodinamičku slobodnu energiju vezivanja ΔG. Razne metode pojednostavljuju ove efekte i aproksimiraju snagu interakcije, definišući posebne skor-funkcije. Najčešće korišćeni parametri su [67]: geometrijska komplementarnost, vodonične veze, kontaktna površina, intermolekulsko i intramolekulsko poklapanje, kontakti para aminokiselina, elektrostatičke interakcije, energija solvatacije, rezidui na aktivnom mestu i slobodna energija.

Tabela 2.1. Lista programa za doking. Navedeni su najčešći alati za doking u procesu virtuelnog skrininga, zajedno sa implementiranom vrstom pretraživanja i skor funkcijom.

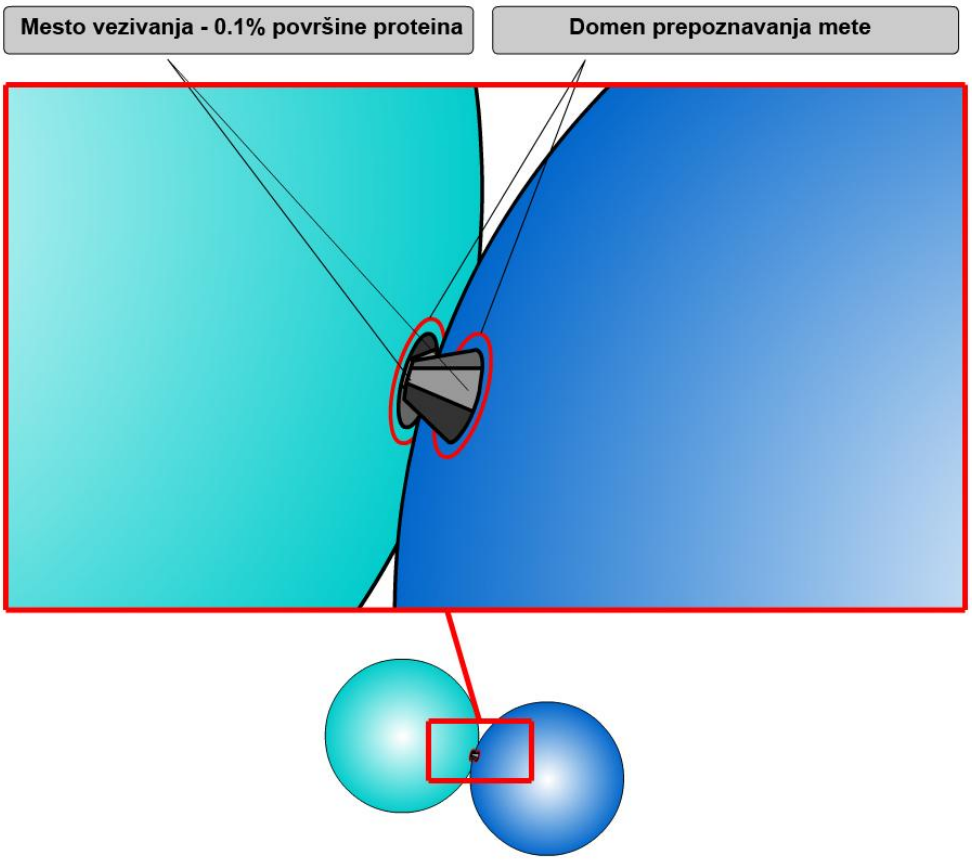
Metoda	**Pretraživanje**	**Skor-funkcija**
DOCK [91, 92]	Fragmentno	Polje sila: AMBER
FlexX [85]	Fragmentno	Empirijska: LUDI
SLIDE [93]	Konformaciono	Empirijska: ChemScore
FRED [94]	Konformaciono	Gausijan / empirijska
GOLD [95]	Genetski algoritam	Empirijska
GLIDE [96]	Iscrpno pretraživanje	Empirijska: ChemScore
AutoDock [97]	Genetski algoritam	Polje sila: AMBER
LigandFit [98]	Monte Karlo	Empirijska
ICM [99]	Pseudo-Brunovsko	Konsenzus / empirijska:

		ECEPP/3
QXP [100]	Monte Karlo	Polje sila: AMBER
AMBER [101]	Molekularna dinamika	Polje sila: AMBER
CHARMM [78]	Molekularna dinamika	Polje sila: CHARMM
MCDOCK [102]	Simulirano kaljenje	Polje sila: CHARMM
Prodock [103]	Monte Karlo	Polja sila: AMBER, ECEPP/3
DIVALI [104]	Genetski algoritam	Polje sila: AMBER
DARWIN [105]	Genetski algoritam	Polje sila: CHARMM
ADAM [106]	Fragmentno	Polje sila: AMBER
Hammerhead [107]	Fragmentno	Empirijska
FLOG [108]	Zasnovan na bazi	Polje sila
EUDOC [109]	Zasnovan na bazi	Polje sila: AMBER
SANDOCK [110]	Konformaciono	Empirijske
FTDOCK [111]	Konformaciono	Polje sila
DockIt [112]	Geometrijska rastojanja	Zasnovane na znanju: PMF
PRO-LEADS [113]	Tabu pretraživanje	Empirijska: ChemScore
Affinity [114, 115]	MC Simulirano kaljenje	Empirijska

2.1.3. Problem neusaglašenosti modela sa realnim rezultatima

Ako se proteini posmatraju kao sfere radijusa 18Å (tipična veličina malih proteina) i ako za kontakt nisu bitne orijentacija sfere, prema formuli Smolučovskog [116] broj kontakata ograničenih difuzijom iznosi ~$7x10^9$ $M^{-1}s^{-1}$.

Pre nego što se formiraju hemijske veze između molekula, potrebno je da su interreagujući molekulski regioni pozicionirani dovoljno blizu (na distanci od oko 2Å) i da su odgovarajući atomi pogodno orijentisani, jer su sile privlačenja u procesu prepoznavanja i vezivanja slabe nekovalentne sile. To znači da je mesto vezivanja na proteinu samo mali procenat njegove površine, oko 0.1% (slika 2.3).



Slika 2.3. Aktivno mesto vezivanja na proteinu predstavlja samo 0.1% njegove površine.

Uzimajući u obzir ovo ograničenje, vrednost broja kontakata ograničenih difuzijom prema trodimenzionalnom modelu „slučajne difuzije“, po formuli Smolučovskog, iznosi ~10^6 $M^{-1}s^{-1}$ za protein-ligand interakciju i ~10^3 $M^{-1}s^{-1}$ za protein-protein interakciju. Shodno tome, Nortrup i Erikson su zapazili da se protein-protein kontakti, kao i brzine biohemijskih reakcija, javljaju od 10^3 do 10^4 puta češće u realnim sistemima, nego što je predviđeno ovim modelom [117]. Uzimajući u obzir češće kontakte između proteina, kao i reakcije između antitela i antigena proteina, ovi autori su predložili da realna vrednost kontakata za kreiranje hemijskih veza iznosi ~10^6 $M^{-1}s^{-1}$ [117]. Neke reakcije se mnogo brže odigravaju nego što to predviđa trodimenzionalni model, npr. brzina katalitičke reakcije superoskid-dismutaze iznosi 10^9 $M^{-1}s^{-1}$ iako je aktivan region vezivanja samo ~0.1% od cele površine (slika 2.3) [118].

2.2. Dalekodosežne međumolekulske interakcije

Da bi se rešio problem neusaglašenosti procenjenih vrednosti brzine protein-protein kontakata teorijskih modela i realnih vrednosti eksperimentalnih rezultata, predloženo je nekoliko modela od kojih su neki: smanjenje dimenzije difuzije [119-121], kontakti proteina zasnovani na hidrodinamičkom upravljanju [122] itd. Jedno od najprihvaćenijih rešenja je model Herberta Freliha [123-125] koji je razvijen 60-tih godina prošlog veka i zasniva se na hipotezi o dalekodosežnim međumolekulskim interakcijama u biološkim sistemima koje deluju na rastojanjima do 1000Å i koje se zasnivaju na rezonantnoj komunikaciji između interreagujućih molekula. Frelih je predložio da su biološki makromolekuli sposobni da pobude dipolne vibracije. Na osnovu ovih vibracija, koje se prostiru na rastojanjima od 100Å do 1000Å, interakcija između makromolekula u polarnoj sredini koja uključuje vodu i lipide, dovodi do pojave dalekodosežnih frekventno-selektivnih privlačnih sila, koje su efikasne na daljinama većim od jedne linearne dimenzije interreagujućih makromolekula (1000Å). Između dva molekula, koji u svojim oscilacijama sadrže istu frekventnu komponentu, dolazi do rezonantne interakcije koja rezultuje specifičnom privlačnom silom između ta dva molekula. Takođe je moguće da biološki makromolekuli privuku male molekule na većim rastojanjima indukujući njihove „pasivne “ oscilacije.

Kao posledica komplementarnosti frekvenci oscilovanja (rezonantnoj komunikaciji) interreagujućih molekula, broj korisnih sudara je povećan u poređenju sa slučajnim sudarima.

Za uspešnu analizu interakcija bioloških molekula potrebno je koristiti oba modela. Objedinjavanjem Fišerove i Frelihove hipoteze došlo se do modela međumolekulskih interakcija koji podrazumeva dva osnovna koraka:

(i) Međusobno prepoznavanje i selektivno privlačenje interreagujućih molekula na rastojanjima većim od 2Å koje ne zavisi od strukture.
(ii) Njihovo međusobno vezivanje slabim nekovalentnim silama koje je uslovljeno strukturnom kompatibilnošću molekula.

Ovo praktično znači da ako bi se lek dizajnirao samo na osnovu strukture ciljnog molekula prema principu kratkodosežnih interakcija, ali bez informacije o „adresi“ koju

daje dalekodosežno prepoznavanje mete, lek bi bio neefikasan. Analogija ovom principu je pecanje: bez mamca na udici, bez obzira kako je udica napravljena, nema uspešnog ulova.

2.2.1. Fizička osnova

U fiziološkim uslovima proteini imaju osobine vibrirajućih, dinamičkih struktura. Mnoštvo trenutnih tehnika (nuklearna magnetna rezonanca, rengdenska strukturna analiza, florescentna depolarizacija, infracrvena spektroskopija, Ramanovo lasersko rasejanje) su pokazali da proteini i njihovi sastavni delovi podležu konformacijskim pomeranjima u vremenskim periodima od femtosekunde do nekoliko minuta. Promene u proteinskoj konformaciji u vremenskom periodu reda veličine nanosekunde, su pobuđene preraspodelom naelektrisanja, kao što su dipolne oscilacije u hidrofobnim regionima proteina [125]. Frelih je sugerisao da bi grupa proteina povezana u zajedničkom gradijentnom polju napona, kao što je membrana polimera citoskeleta, oscilovala koherentno u periodama od nanosekundi, ako bi postojao izvor energije kao što je npr. biohemijska ATP [124]. Frelihov model koherencije objašnjava efekat dalekodosežne saradnje u kojem proteini i nukleoditne kiseline u biološkim sistemima komuniciraju. Glavna konstatacija Frelihove teorije je da promenljiv izvor energije u sistemu nelinearno vezanih dipola može dovesti do koherentnog pobuđivanja stanja pojedinačnih vibracija, ako dovedena energija pređe određeni prag. Oscilacije su reda veličine 10^9 do 10^{11} Hz i pojedinačne su jer su ostale u stanju temperaturne ravnoteže. Ove koherentna pobuđenja omogućavaju dalekodosežnu saradnju između interreagujućih bioloških molekula.

Frelih je takođe sugerisao mogućnost pobuđenih talasa zbog superprovodljivosti biomolekula. Da bi se dobila superprovodljivost u makromolekulima potrebna je neka vrsta kohezivne energije. U konvencionalnim superprovodnicima to se postiže elektron-elektron interakcijom izazvanom fononom. Litl je predložio sličan mehanizam, ali za makromolekule [126]. On je posmatrao protein kao dugačak lanac tj. „kičmu“ sa elektronima koji imaju razna stanja i mogu formirati sistem prenosa i niza sporednih lanaca povezanih na kičmu. Zatim je pokazao da, iako je „kičma“ na početku izolator, zbog prisustva zakačenih sporednih lanaca, povećava se privlačenje između elektrona

do tačke gde je energija dovoljna da se „kičma“ pretvori iz stanja izolatora ili poluprovodnika u superprovodnik. Stanje superprovodnika generisano na ovaj način bi se javljalo čak i na temperaturama iznad sobne temperature.

Protok elektrona kroz „kičmu“ izaziva polarizaciju sporednih lanaca, što omogućava kratkotrajnu interakciju sa drugim polaronima i puni kiselinske ostatke proteina naelektrisanjem. Ove kratkotrajne privlačne i odbijajuće interakcije rezultuju dinamikom proteina, koja zavisi od fizičkih karakteristika sporednih lanaca „kičme“. Ta dinamika proteina generisana protokom elektrona u „kičmi“ je definisana primarnom strukturom proteina i prenosi se susednim molekulima vode. Na ovaj način informacija kodirana u primarnoj strukturi proteina može biti prenesena i raširena kroz vodu preko Frelihovih oscilacija. Takođe, Frelihove oscilacije izazivaju pasivne oscilacije polidelokalizovanih valentnih elektrona u susednim malim molekulima, što dovodi do selektivne interakcije između malih molekula i makromolekula [127].

2.2.2. EIIP/AQVN koncept za male molekule

Kao osnovni fizički parametri bioloških molekula koji određuju osobine njihovih dalekodosežnih međumolekulskih veza, uzimaju se broj valentnih elektrona i potencijal elektron-jon interakcije (eng. *Electron-Ion Interaction Potential*, ***EIIP***) [127]. Sve međumolekulske interakcije su definisane raspodelom i energijom valentnih elektrona u molekulima, pošto oni učestvuju u određivanju bitnih svojstava kao što su: dipolni i kvadropolni momenat, jonizacioni potencijal, elektrofilnost itd. Pokazano je [128] da se EIIP za organske molekule može izračunati na osnovu jednostavne formule, izvedene iz “opšteg modela pseudopotencijala” [129-131]:

$$W = 0.25\frac{Z^* \sin(1.04\pi Z^*)}{2\pi} \tag{2.1}$$

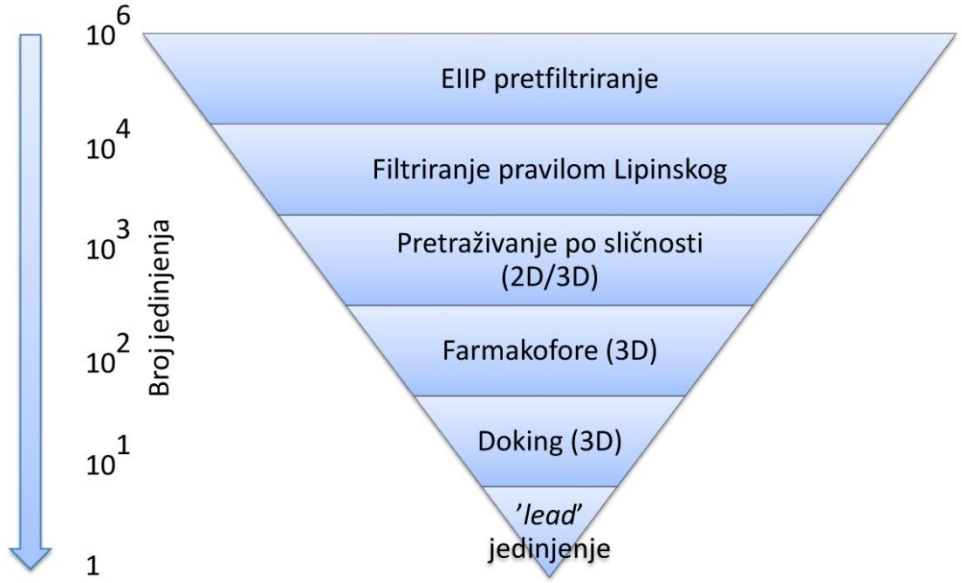
gde je Z^* srednji kvazivalentni broj (eng. *Average Quasivalence Number*, ***AQVN***) određen formulom:

$$Z^* = \frac{1}{N}\sum_{i=1}^{m} n_i Z_i \tag{2.2}$$

gde je Z_i valentni broj atoma koji je i-ta komponenta molekula, n_i broj atoma i-te komponente, m broj komponenti molekula, N ukupan broj atoma u molekulu. EIIP

vrednost molekula izračunata na osnovu jednačina (2.1) i (2.2) ima jedinicu mere Rydberg (Ry). Visoka korelacija je pokazana između EIIP i AQVN vrednosti za organske molekule i njihovih bioloških aktivnosti. Neke od tih osobina su: mutagenost, karcigenost, toksičnost, aktivnost antibiotika i citostatika itd. [127, 128, 132-136].

Objedinjavanjem principa kratkodosežnih i dalekodosežnih međumolekulskih interakcija u metodi virtuelnog skrininga, poboljšava se i ubrzava proces otkrivanja leka, i omogućava analiza većeg broja jedinjenja. VS zasnovan na AQVN/EIIP vrednostima se sastoji od koraka predfiltriranja po AQVN/EIIP vrednostima jedinjenja, pre standardnog strukturno zasnovanog skrininga molekulskih biblioteka.



Slika 2.4. Shematski prikaz toka protokola virtuelnog skrininga u procesu otkrivanja leka, zasnovan na objedinjenom pristupu kratkodosežnih i dalekodosežnih interakcija.

3. Materijal i metode

3.1. Proteinske i nukleotidne baze podataka

3.1.1. Nukleotidne baze

Jedna od prvih baza nukleotidnih sekvenci je formirana 1979. godine u Nacionalnoj Laboratoriji Los Alamos (*Los Alamos National Laboratory - LANL*), a 1982. godine je pretvorena u javnu bazu pod nazivom ***GenBank*** [137]. Dok je na početku sadržala oko 2000 sekvenci, u februaru 2013. sadržala je preko 162 miliona sekvenci iz više od 100 hiljada različitih organizama. Bazu održava i vlasnik je NCBI [138].

Pored GenBank baze, 1980. godine je u Evropskoj laboratoriji za molekularnu biologiju osnovana ***EMBL*** biblioteka nukleotidnih sekvenci [139]. Bazu održava udruženje sastavljeno od 20 članica država iz Evrope, sa Izraelom i Australijom.

U Japanu u Nacionalnom institutu za genetiku, 1984. godine osnovana je ***DDBJ*** baza DNK sekvenci [140]. GenBank, EMBL i DDBJ su 1987. godine oformile Internacionalnu kolaboraciju baza nukleotidnih sekvenci INSDC [141] koja funkcioniše tako što baze svakodnevno međusobno razmenjuju podatke [142].

3.1.2. Proteinske baze

Prva baza proteinskih sekvenci, pod nazivom ***PIR*** baza (eng. *Protein Information Resource*), je ustanovljena 1984. godine od strane Američke nacionalne biomedicinske fondacije iz Vašingtona (NBRF) [143]. Prethodno je NBRF skupila prvu kolekciju sekvenci od 1965-1978. godine u knjizi „Atlas proteinskih sekvenci i struktura“ autorke Margaret O. Dazhoff [144].

Pored PIR baze, 1986. godine je kreirana ***Swiss-Prot*** baza proteinskih sekvenci, od strane Švajcarskog instituta za bioinformatiku SIB [145] i Evropskog bioinformatičkog instituta [146], a zajedno sa njom i ***TrEMBL*** (*Translated EMBL*

Nucleotide Sequence Data Library) baza prevedenih EMBL nukleotidnih sekvenci koje nisu u Swiss-prot bazi.

Godine 2002. PIR je sa partnerima EBI i SIB dobio sredstva od Američkog nacionalnog instituta za zdravlje NIH [147] od kojih su kreirali ***UniProt*** - jedinstvenu svetsku bazu proteinskih sekvenci i funkcija [148].

GenomeNet je mreža baza sekvenci, osnovana 1991. godine u Kjoto Univerzitetu u Japanu [149]. Čini ga mreža baza i servisa za istraživanja u biomedicini. Sastoji se iz pet kategorija baza: lokalna baza **KEGG** (*Kyoto Encyclopedia of Genes and Genomes*), ostale lokalne baze, pretraživanje ostalih javnih baza na internetu, veze ka drugim javnim bazama (PubChem i INSDC) i PubMed baze.

3.1.3. Servisi za pristup bazama sekvenci

Pored samih baza sekvenci, veoma bitan je i pristup bazama putem interneta. Zato su razvijeni razni veb servisi koji omogućavaju razne upite na bazama sekvenci, kao i samu analizu sekvenci u bazi razvijenim bioinformatičkim alatima.

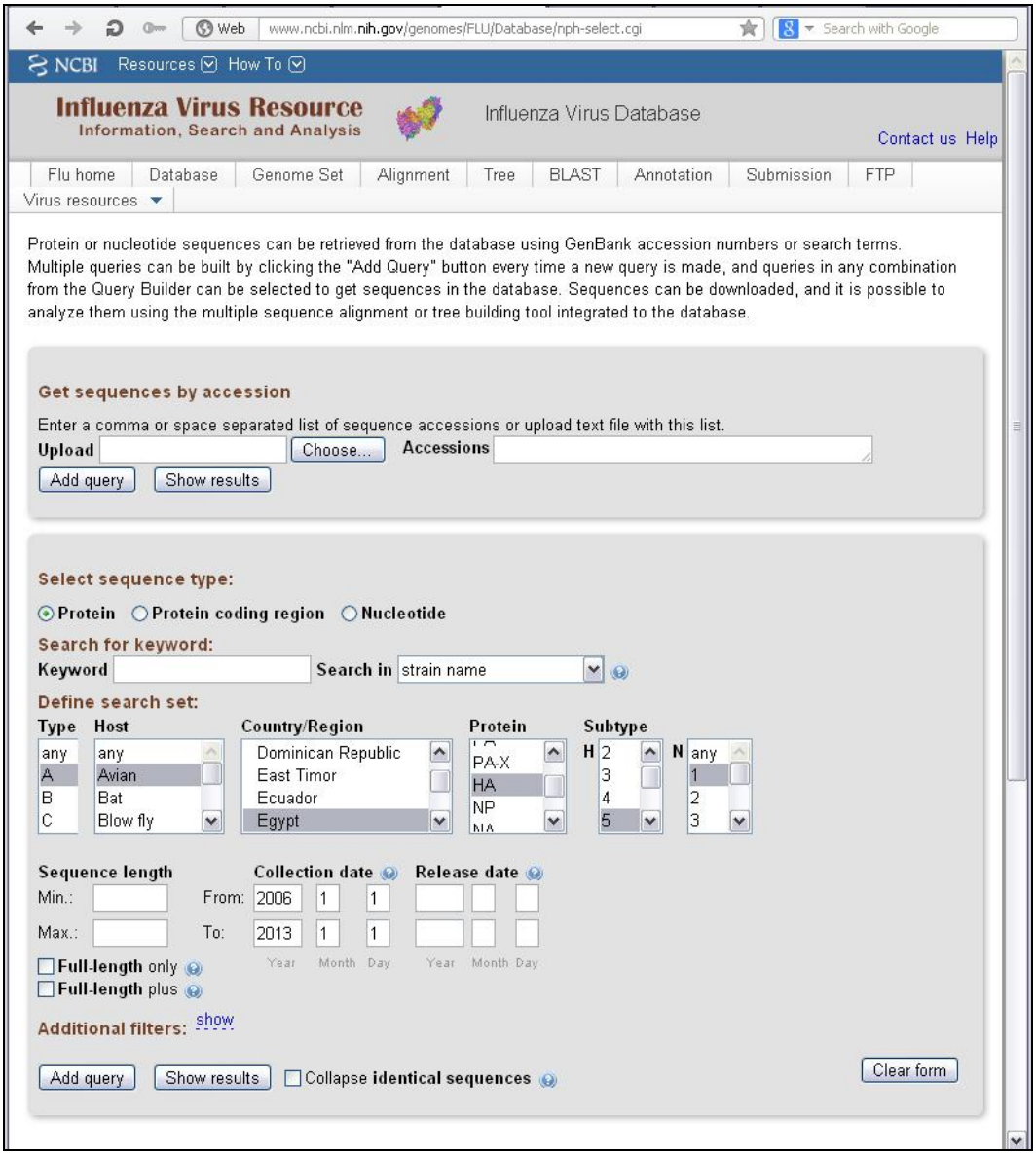
Među prvim servisima je razvijen program **ENTREZ** u NBCI centru [150]. Poseduje jednostavan interfejs, fleksibilno i precizno pretraživanje nukleotidnih sekvenci, genoma, gena, proteina, malih molekula, ćelijskih puteva i ekspresija gena. ENTREZ sistem čini više od 40 baza molekula i literature, među kojima su: GenBank, EMBL, DDBJ, INSDC, EST (*Expressed Sequence Tag*), GSS (*Genome Survey Sequence*), BioSystems, Gene, Genome, GEO (*Gene Expression Omnibus*), PIR, Swiss-Prot, PDB (*Protein Data Bank*), PubChem, PubMed, Taxonomy, NCBI Bookshelf, NLM Catalog, (*National Library of Medicine in Washington DC*).

Drugi, takođe često korišćen sistem za pristup bazama, je **DBGET** [151], koji pristupa mreži baza *GenomeNet*.

Projekat sekvenciranja genoma influence [152], osnovan od strane Američkog nacionalnog instituta za alergije i infektivne bolesti (NIAID) [153], ima za cilj širenje znanja o influenci, posebno kako se virusi gripa razvijaju, šire i uzrokuju bolesti, sa svrhom razvijanja novih vakcina, terapija, dijagnostike, praćenja evolucije influence, pojava pandemija gripa i ublažavanja posledica epidemija. NIAID pruža podatke o kompletnim sekvencama virusa influenca iz GenBank baze, podatke istraživanja o

virusima kroz Bazu istraživanja influence [154], zajedno sa alatima za analizu podataka kroz jedinstven internet servis.

NCBI je razvio servis **Resursi Virusa Influence** [155] koji pruža javnu uslugu pretraživanja i analize podataka iz NIAID i GenBank baza, vezanih za viruse influence, kroz razne alate za poravnavanje, klasterisanje, filogenetsku analizu i anotaciju sekvenci. Baza influenca virusa je dostupna na URL adresi (slika 3.1): http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi.



Slika 3.1. Internet stranica servisa NCBI Influenca Virus za pristup bazi virusa influence.

3.1.4. Formati zapisa sekvenci

Sekvence se zapisuju u tekstualnim datotekama standardnim ASCII karakterima. Za zapis same sekvence, nukleotidne ili proteinske, koristi se jednoslovni kôd (tabela 3.6). Različiti programi koriste različite formate zapisa sekvenci, koji na različite načine označavaju tipove informacija o sekvenci. Najčešći formati su: GenBank, EMBL, SwissProt, FASTA, NBRF (PIR). Posebno, za potrebe EIIP/ISM platforme razvijen je format SEQ za zapis sekvenci.

3.1.4.1. GenBank format

Zapis sekvenci u NCBI proteinskim i nukleotidnim bazama sadrži informacije o: referencama, funkciji, lokacijama RNK i kodirajućih regiona i poziciji bitnih mutacija. Informacije su organizovane pomoću polja, tako da je početak svakog polja identifikator koji označava vrstu informacije (tabela 3.1). Svako polje počinje u novom redu. Polje FEATURES sadrži podpolja. Zapis same sekvence se nalazi između simbola ORIGIN i kraja zapisa „//“. Poseban simbol za kraj zapisa jedne sekvence „//“, omogućava zapis više sekvenci i jednu datoteku, gde se svaka sekvenca zapisuje jedna ispod druge.

Primer zapisa humanog insulina u GenBank formatu:

```
LOCUS       BT006808  333 bp    mRNA    linear   PRI 13-MAY-2003
DEFINITION  Homo sapiens insulin mRNA, complete cds.
ACCESSION   BT006808
VERSION     BT006808.1  GI:30582454
KEYWORDS    FLI_CDNA.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 333)
  AUTHORS   Kalnine,N., Chen,X., Rolfs,A., Halleck,A., Hines,L., Eisenstein,S.,
            Koundinya,M., Raphael,J., Moreira,D., Kelley,T., LaBaer,J., Lin,Y.,
            Phelan,M. and Farmer,A.
  TITLE     Cloning of human full-length CDSs in BD Creator(TM) System Donor
            vector
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 333)
  AUTHORS   Kalnine,N., Chen,X., Rolfs,A., Halleck,A., Hines,L., Eisenstein,S.,
            Koundinya,M., Raphael,J., Moreira,D., Kelley,T., LaBaer,J., Lin,Y.,
            Phelan,M. and Farmer,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (13-MAY-2003) BD Biosciences Clontech, 1020 East Meadow
            Circle, Palo Alto, CA 94303, USA
COMMENT     This CDS clone is a part of a collection of human full length
            expression clones generated by BD Biosciences Clontech and the
            Harvard Institute of Proteomics.
```

```
FEATURES             Location/Qualifiers
     source          1..333
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /clone="GH00103X1.0"
                     /clone_lib="BD Creator(TM) CDS Library derived from MGC
                     collection"
                     /lab_host="DH5alpha T1 resistant"
                     /note="Vector: pDNR-Dual"
     CDS             1..333
                     /codon_start=1
                     /product="insulin"
                     /protein_id="AAP35454.1"
                     /db_xref="GI:30582455"
                     /translation="MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCG
                     ERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSL
                     YQLENYCN"
ORIGIN
        1 atggccctgt ggatgcgcct cctgcccctg ctggcgctgc tggccctctg gggacctgac
       61 ccagccgcag cctttgtgaa ccaacacctg tgcggctcac acctggtgga agctctctac
      121 ctagtgtgcg gggaacgagg cttcttctac acacccaaga cccgccggga ggcagaggac
      181 ctgcaggtgg ggcaggtgga gctgggcggg ggccctggtg caggcagcct gcagcccttg
      241 gccctggagg ggtccctgca gaagcgtggc attgtggaac aatgctgtac cagcatctgc
      301 tccctctacc agctggagaa ctactgcaac tag
//
```

Tabela 3.1. Lista bitnih identifikatora u GenBank formatu zapisa sekvenci.

Identifikator	**Opis**
LOCUS	izvor, klasa organizma, dužina i tip sekvence
DEFINITION	opis sekvence
ACCESSION	identifikacioni broj
KEYWORDS	ključne reči za vezu sa drugim bazama
ORGANISM	izvorni organizam
COMMENT	opis biološke funkcije
FEATURES	informacije vezane za posebne pozicije ili interval pozicija
CDS	/translation="kodirana proteinska sekvenca"
MUTATION	pozicija mutacije i nova baza
ORIGIN	početak zapisa DNK sekvence
//	kraj zapisa sekvence

3.1.4.2. EMBL format

EMBL format je, slično GenBank formatu, organizovan pomoću polja koja počinju identifikatorom. Identifikatori su dvoslovne skraćenice opisa i mogu imati više mogućih zapisa (tabela 3.2).

Tabela 3.2. Lista bitnih identifikatora u EMBL formatu zapisa sekvenci, sa opisom njihovih značenja.

Identifikator	**Opis**
ID	identifikacioni broj
DE	opis
AC	pristupni identifikacioni broj (eng. *accession number*)
DT	datum unosa
KW	ključne reči za vezu sa drugim bazama
OS	izvorni organizam
CC	biološka funkcija
FH,FT	informacije vezane za posebne pozicije ili interval pozicija
CDS	/translation="kodirana proteinska sekvenca"
MUTATION	pozicija mutacije i nova baza
SQ	početak zapisa DNK sekvence
//	kraj zapisa sekvence

3.1.4.3. SwissProt format

Format zapisa proteina u SwissProt bazi je isti kao i EMBL format, s tim što poseduje više informacija i vrsta polja.

3.1.4.4. FASTA format

FASTA format za zapis sekvence je vrlo jednostavan i sastoji se iz dva dela. U prvom redu je komentar o sekvenci koji počinje specijalnim simbolom „>“. U narednim redovima je sam zapis sekvence. Komentar o sekvenci u prvom redu ima svoj format koji zavisi od baze i verzije, čime se omogućava zapis više vrsta informacija o sekvenci. Kao opcioni simbol za kraj zapisa sekvence je „*“, ali ne mora biti prisutan, jer simbol „>“ na početku reda, kao oznaka početka zapisa sledeće sekvence, služi kao separator.

Najčešći format komentara je:

>*entryName desc*

gde su: *entryName* – naziv sekvence, *desc* – opis sekvence.

U Uniprot bazi format komentara je:

>*ID*|*entryName desc*

gde su: *ID* – identifikator, *entryName* - naziv i *desc* – opis sekvence.

Pored najčešćih, postoje i drugi formati komentara:

>*accession*|*name desc*

gde su: *accession* - pristupni identifikator, *name* - naziv i *desc* - opis sekvence.

>*source*|*accession*|*name_locus desc*

gde su: n*ame_locus* naziv ili tip sekvence, *source* - oznaka izvorne baze koja može biti: „sp“ za SwissProt i „tr“ za Trembl bazu.

>sp|*ID*|*entryName desc*

gde je simbol „sp“ oznaka da je izvor SwissProt baza.

Primer zapisa proteina u FASTA formatu:

```
>ABM92273|A/Egypt/0636-NAMRU3/2007
DQICIGYHANNSTEQVDTIMEKNVTVTHAQDILEKTHNGKLCNLNGVKPLILRDCSVAGW
LLGNPMCDEFLNVPEWSYIVEKINPANDLCYPGNFNDYEELKHLLSRINHFEKIQIIPKN
SWSDHEASGVSSACPYQGRSSFFRNVVWLTKKDNAYPTIKRSYNNTNQEDLLVLWGIHHP
NDAAEQTRLYQNPTTYISVGTSTLNQRLVPKIAARSKVNGQSGRMEFFWTILKSNDAINF
ESNGNFIAPENAYKIVKKGDSTIMKSELEYGNCNTKCQTPIGAINSSMPFHNIHPLTIGE
CPKYVKSNRLVLATGLRNSPQGERRRKKR
```

3.1.4.5. SEQ format

Za potrebe EIIP/ISM platforme, razvijen je lokalni format zapisa pojedinih proteinskih i nukleotidnih sekvenci. Datoteka za čuvanje sekvence SEQ formata ima ekstenziju *.seq.

SEQ format ima sledeći oblik:

- Prvi red označava vrstu sekvence (Protein, DNK ili RNK).
- Drugi red je opis sekvence.

- Od trećeg reda na dalje je zapis same sekvence sa mogućim znacima belina zbog čitljivijeg zapisa, a koji ne utiču na tumačenje sekvence.

Primer zapisa proteina u SEQ formatu:

```
Protein
ABM92273|A/Egypt/0636-NAMRU3/2007
DQICIGYHANNSTEQVDTIMEKNVTVTHAQDILEKTHNGKLCNLNGVKPLILRDCSVAGW
LLGNPMCDEFLNVPEWSYIVEKINPANDLCYPGNFNDYEELKHLLSRINHFEKIQIIPKN
SWSDHEASGVSSACPYQGRSSFFRNVVWLTKKDNAYPTIKRSYNNTNQEDLLVLWGIHHP
NDAAEQTRLYQNPTTYISVGTSTLNQRLVPKIAARSKVNGQSGRMEFFWTILKSNDAINF
ESNGNFIAPENAYKIVKKGDSTIMKSELEYGNCNTKCQTPIGAINSSMPFHNIHPLTIGE
CPKYVKSNRLVLATGLRNSPQGERRRKKR
```

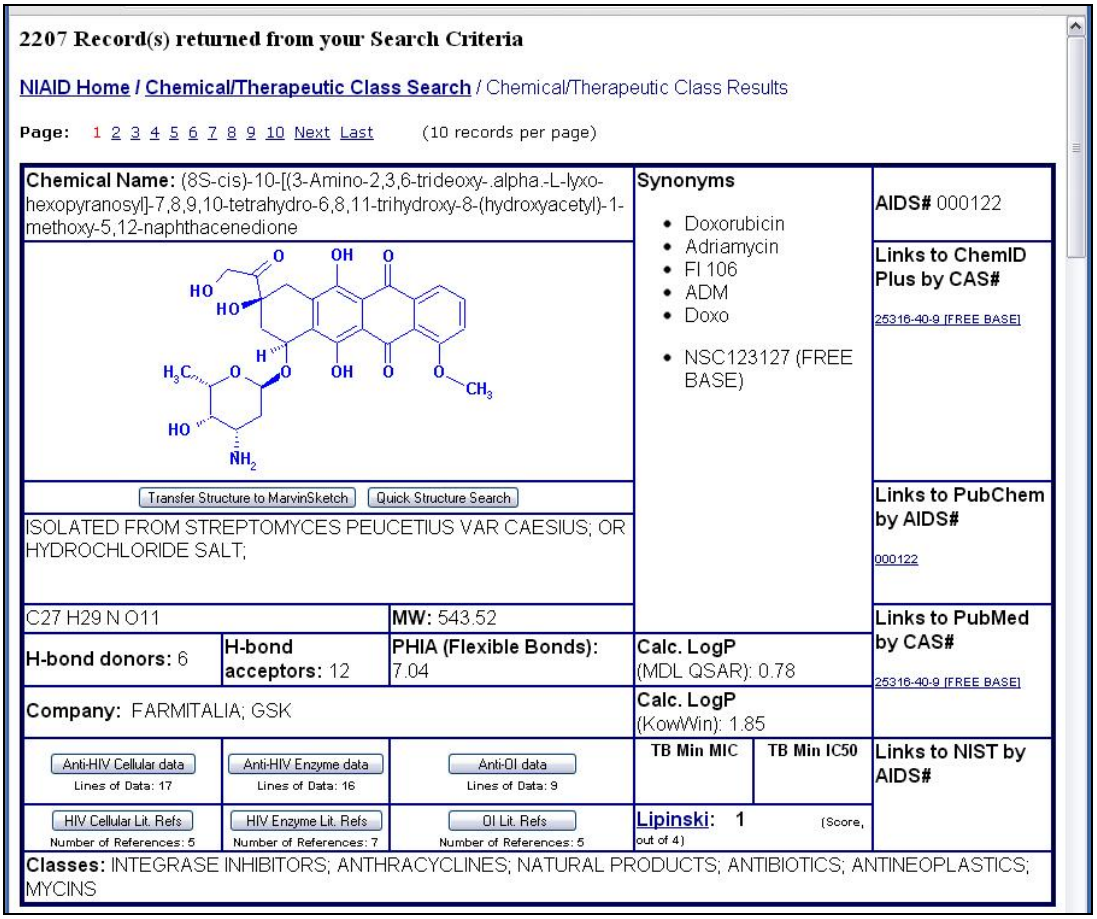
3.2. Molekulske biblioteke

ChemDB baza

Američki nacionalni instituta za alergije i infektivne bolesti (NIAID) je 1989. godine razvio i predstavio ChemDB bazu malih molekula [156], koja sadrži testirane terapeutike za HIV, tuberkulozu i ostale oportunističke infekcije kao što su: SIV, FIV, humani citomegalovirus - HCMV, herpes simpleks virus 1 i 2, hepatitis A,B i C, kandida, Epšten-bar virus, plasmodium, toksoplazma, mikrosporidia, itd. ChemDB je alat za prikupljanje i arhiviranje pretkliničkih podataka o jedinjenjima kao potencijalnim terapeuticima protiv HIV-a i sličnih oportunističkih infekcija. Podaci o strukturi i funkciji testiranih jedinjenja iz objavljene literature se konstantno prikupljaju i organizovani su prema hemijskim osobinama i biološkim aktivnostima.

Pretraživanje jedinjenja se može izvršiti prema: (i) strukturi po poljima: *Chemical Name, Company, Chemical Class, AIDS#, Molecular Formula, Molecular Weight*, ili hemijskim karakteristikama koristeći zapis u SMILES formatu, (ii) aktivnosti na određene patogene ili enzime, (iii) tekstualnim informacijama u objavljenim člancima (npr. godina, autor, naslov, časopis).

Rezultat pretrage jedinjenja je u formatu HTML, što ga čini pogodnim za vizuelan prikaz u internet pregledaču, ali je za izvođenje informacija o jedinjenjima potrebno parsirati (sintaksno analizirati) html datoteku (slika 3.2).



Slika 3.2. Internet stranica rezultata pretrage na servisu ChemDB.

Za svako jedinjenje, rezultat pretrage sadrži sliku strukture i informacije koje su raspoređene po poljima. Neke od bitnih informacija su: naziv jedinjenja (*Chemical Name*), sinonimi, opis, molekularna formula (*Formula*), molekulska težina (*MW*), naziv kompanije (*Company*), identifikator u bazi (*AIDS#*), identifikatori veza sa drugim bazama, klase kojima pripada jedinjenje (*Classes*).

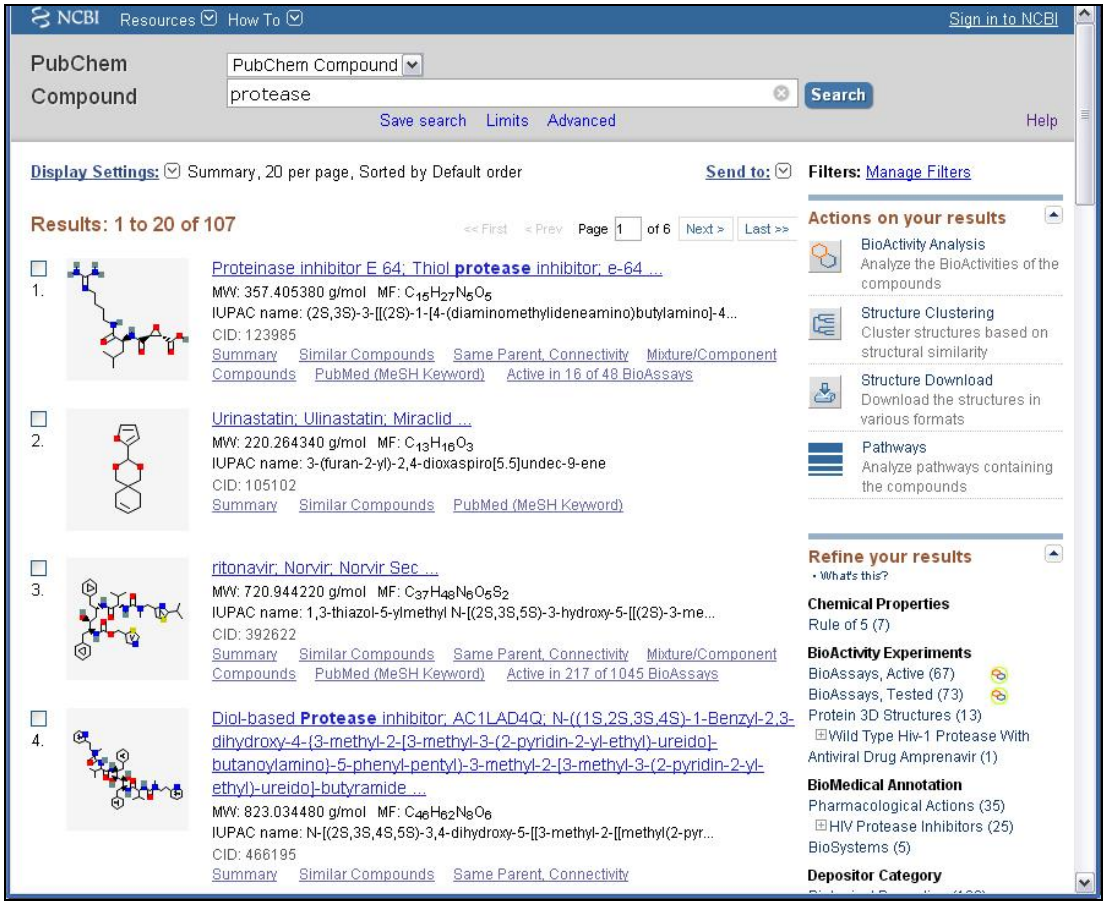
PubChem baza

PubChem je baza informacija o biološkim funkcijama malih molekula [157]. Baza je razvijena 2004. godine u institutu NCBI i sastoji se od tri podbaze: (i) PubChem Supstance (sadrži oko 119 miliona jedinjenja, kompleksa i nekarakterizovanih supstanci), (ii) PubChem Jedinjenja (sastoji se od oko 47 miliona karakterizovanih jedinjenja), (iii) PubChem BioAssay (sadrži oko 717 hiljada bioeseja sa više miliona

bioloških funkcija dobijenih iz raznih programa za skrining) [3]. PubChem baza je povezana na *Entrez* bazu koja sadrži biološke informacije o 3D strukturama jedinjenja.

Pored brze pretrage jedinjenja po hemijskoj strukturi, PubChem servis sadrži i razne alate za klasifikaciju, klasterisanje po strukturama i 3D konformaciju.

Kao i kod ChemDB baze, rezultat pretrage jedinjenja je u formatu HTML, pa je za izvođenje informacija o jedinjenjima (tabela 3.3) potrebno parsirati HTML datoteke (slika 3.3).



Slika 3.3. Internet stranica rezultata pretrage na servisu PubChem.

Internet servis za pristup PubChem bazi omogućava preuzimanje jedinjenja iz baze u obliku datoteka u ASN, XML ili SDF formatu, ali samo preko pretrage po identifikatorima. Pored toga je moguć prenos cele baze preko FTP protokola, preuzimanjem pojedinačnih datoteka od po 25000 jedinjenja.

Tabela 3.3. Nazivi polja u rezultatu pretrage na servisu PubChem, koja sadrže informacije o jedinjenjima.

Identifikator	**Opis**
UID	identifikator u trenutnoj bazi
SID	identifikator u bazi supstanci
CID (Compound_ID)	identifikator u bazi jedinjenja
BAID	identifikator u bazi BioAssays
IUPAC name	naziv jedinjenja
MW	molekulska težina
Source	informacije o izvoru

ASINEX baza

Asinex je kompanija locirana u Moskvi koja pruža usluge dizajna lekova, i specijalizovana je za kreiranje novih jedinjenja, modli i biblioteka malih molekula [158]. Pored komercijalnih servisa koje kompanija pruža, Asinex omogućava slobodan pristup bibliotekama jedinjenja. ASINEX biblioteke sadrže informacije o 600 hiljada jedinjenja, zajedno sa njihovim 3D-strukturama, koje su podeljene u nekoliko posebnih biblioteka [159]:

- Zlatna i Platinumska kolekcija je kreirana 1994. godine i sadrži preko 300 hiljada jedinjenja koja imaju visok stepen pogodnosti za lekove i filtrirana je prema pravilu Lipinskog [160].
- Elitna i Sinergetska biblioteka je razvijena kao rezultat raznih programa za dizajn lekova i sadrži preko 100 hiljada jedinjenja
- BioDizajn bibilioteka sadrži preko 80 hiljada jedinjenja i 600 sintetizovanih modli baziranih na strukturnim osobinama alkaloida.
- Minimalističke biblioteke su bazirane na prirodnim proizvodima i sačinjene od 17 hiljada jedinjenja i 125 sintetizovanih skeleta modli zasnovanih na analizi prirodnih alkaloida, lekova i komercijalnih biblioteka jedinjenja.

Sve Asineks baze jedinjenja su date u SDF formatu.

3.2.1. Formati zapisa molekulskih jedinjenja

SDF format

SDF format (eng. *Structure Data File*) je razvila firma MDL (*Molecular Design Limited*) sa svrhom skladištenja informacija o hemijskim jedinjenjima [161]. Format zapisa se sastoji iz dva dela:

Prvi deo opisuje molekulsku strukturu i sadrži informacije o atomima, vezama i koordinatama atoma i molekula, i sledećeg je formata:

Naslov

Tabela hemijskih veza

gde polje *Naslov* sadrži naziv jedinjenja, datum, komentar i druge informacije o molekulu, dok polje *Tabela hemijskih veza* sadrži informacije o atomima i njihovim koordinatama i vrsti atomskih veza.

Drugi deo sadrži asocirane podatke o jedinjenju. Podaci seređaju jedan za drugim, a svaki podatak je formata:

Naslov podatka

Podatak

Prazan red

gde je *Naslov podatka* identifikator informacije koji je u jednom redu i počinje simbolom „>“ za kojim ide naziv polja u zagradama „< >“, ili broj polja „DTn“, sa opcionim internim registarskim brojem (tabela 3.4). *Podatak* se može nalaziti u više linija. *Prazan red* je oznaka za kraj podatka, odnosno predstavlja terminator zapisa svakog podatka.

„*$$$$*“ je simbol za oznaku kraja zapisa jednog jedinjenja (terminator), što omogućava zapis više jedinjenja u jednu datoteku.

Primer zapisa vode u SDF formatu:

```
962
  -OEChem-07091322092D

  3  2  0     0  0  0  0  0  0999 V2000
    2.5369   -0.1550    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    3.0739    0.1550    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    2.0000    0.1550    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
```

```
  1  2  1  0  0  0  0
  1  3  1  0  0  0  0
M  END
> <PUBCHEM_COMPOUND_CID>
962

> <PUBCHEM_MOLECULAR_WEIGHT>
18.01528

> <PUBCHEM_EXACT_MASS>
18.010565

> <PUBCHEM_MOLECULAR_FORMULA>
H2O

> <PUBCHEM_IUPAC_INCHI>
InChI=1S/H2O/h1H2

> <PUBCHEM_IUPAC_INCHIKEY>
XLYOFNOQVPJJNP-UHFFFAOYSA-N

> <PUBCHEM_OPENEYE_ISO_SMILES>
O

> <PUBCHEM_IUPAC_NAME>
oxidane

> <PUBCHEM_IUPAC_OPENEYE_NAME>
water

$$$$
```

Tabela 3.4. Značajnija polja u SDF zapisu jedinjenja u PubChem bazi.

Identifikator	**Opis**
PUBCHEM_COMPOUND_CID	identifikator u PubChem bazi supstanci
PUBCHEM_IUPAC_NAME	naziv jedinjenja
PUBCHEM_MOLECULAR_FORMULA	molekularna formula
PUBCHEM_MOLECULAR_WEIGHT	molekulska težina
PUBCHEM_OPENEYE_ISO_SMILES	zapis jedinjenja u SMILES fomatu

SMILES format zapisa molekula

SMILES (eng. *Simplified Molecular Input Line Entry System*) je specijalan format koji omogućava da se dvodimenzionalna struktura ili trodimenzionalni model hemijskog jedinjenja zapiše u linearnoj tekstualnoj formi, zbog jednostavnog zapisa u bazama i korišćenja u programima.

SMILES notacija se sastoji od niza karaktera bez praznih polja. Atomi vodonika mogu biti ubačeni ili izostavljeni iz zapisa.

Osnovna pravila SMILES formata su [162-165]:

1. Atomi su predstavljeni svojim simbolima, i to je jedini slučaj upotrebe slova u zapisu. U aromatizovanim lancima atomi se zapisuju malim slovima. Ako valenca atoma i broj vezanih vodonika nije „najmanji normalan“, atom se mora staviti u zagrade „[ ]“. Ako je valenca vodonika normalna, oznaka za vodonik se izostavlja, inače se iza simbola H zapisuje opcioni broj vezanih atoma vodonika i znak naelektrisanja.

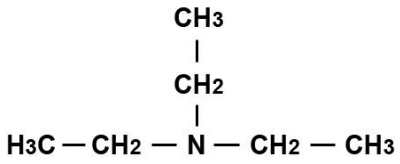
   Primeri: O za vodu (H2O); C za metan (CH4); [H+] za proton.

2. Jednostruke, dvostruke, trostruke i aromatične hemijske veze se predstavljaju simbolima „-“ „=“ „#“ i „:“ redom. Jednostruka veza je podrazumevana i može biti izostavljena.

   Primer: C#N za cijanovodoničnu kiselinu (HCN).

3. Svako grananje je predstavljeno tako što se svaka grana zatvori u zagrade „( )“, koje mogu biti ugnježdene.

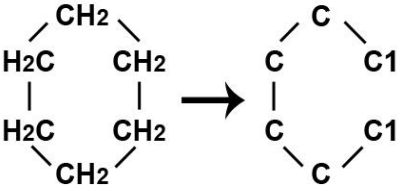
   Primer: CCN(CC)CC za trietilamin.



Slika 3.4. Trietilamin.

4. Ciklične strukture se predstavljaju tako što se rasturi jedna veza prstena, atomi te veze se indeksiraju istim brojem koji je u zapisu odmah iza tih atoma. Za svaki prsten se koriste različiti indeksi.

   Primer: C1CCCCCC1 za cikloheksan.

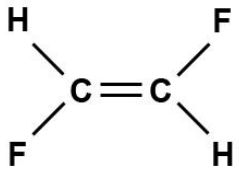


Slika 3.5. Cikloheksan.

5. Nepovezani delovi jedinjenja se zapisuju kao pojedinačne strukture povezana simbolom „.“ .
6. Izotopi se predstavljaju masenim brojem ispred simbola atoma u zagradama „[]“.

   Primer: [12C] za ugljenik-12.
7. Za smer dvostruke veze koriste se simboli „/“ i „\“.

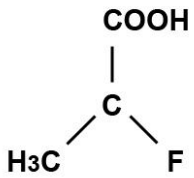
   Primer: F/C=C\F za E-1,2-difloretan.



Slika 3.6. E-1,2-difloretan.

8. Kod označavanja hiralnosti, za oznaku tetraedarnog centra se koristi simbol „@“ ili „@@“ iza oznake atoma. Simbol @ označava da se susediređaju u smeru suprotnom od skazaljke, dok simbol @@ određuje smer skazaljke.

   Primer: N[C@](C)(F)C(=O)O za metil-C,F,karboksi-C.



Slika 3.7. Primer hiralnosti.

Tabela 3.5. Primeri SMILES zapisa nekih jedinjenja.

SMILES zapis	**Naziv jedinjenja**
O=C=O	ugljen dioksid
CC(=O)O	sirćetna kiselina
N[C@@H](C)C(=O)O	alanin
CC(=O)NCCC1=CNc2c1cc(OC)cc2	melatonin

3.3. Furijeova transformacija

Furijeova transformacija je osnovna analitička metoda digitalne obrade signala, koja omogućava predstavljanje signala u frekvencijskom domenu. Furijeovom transformacijom signala dobija se spektar signala.

3.3.1. Furijeova transformacija diskretnog signala

Za diskretan niz kompleksnih brojeva $\{x(n)\}$ Furijeova transformacija je definisana formulom [166]:

$$X(e^{i\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-i\omega n} \tag{3.1}$$

gde je $X(e^{i\omega})$ kontinualna i periodična kompleksna funkcija po ω sa periodom 2π, a ako $\{x(n)\}$ predstavlja signal $X(e^{i\omega})$ se naziva frekvencijski spektar signala.

Inverzna Furijeova transformacija, koja daje članove niza $\{x(n)\}$, je određena Furijeovim integralom:

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{i\omega})e^{i\omega n} d\omega \tag{3.2}$$

Kod Furijeove transformacije realnih nizova važe osobine da su realni deo Furijeove transformacije $X(e^{i\omega})$ i amplitudna karakteristika parne funkcije, a imaginarni deo i fazna karakteristika neparne funkcije [166], pa su u frekvencijskom opsegu $0 \leq \omega \leq \pi$ sadržane sve informacije Furijeove transformacije realnog niza.

3.3.2. Diskretna Furijeova transformacija (DFT)

Diskretna Furijeova transformacija (eng. *Discrete Fourier Transform*, DFT) je nastala diskretizacijom jedne periode Furijeove transformacije (kontinualne funkcije) diskretnog signala, što omogućava obradu diskretnih signala na računaru. Odabiranje u frekvencijskom domenu se izvršava po N ekvidistantnih tačaka u intervalu $[0,2\pi)$, gde su kružne frekvence ω_k date formulom:

$$\omega_k = \frac{2\pi k}{N}, \quad k = 0,\ldots, N-1 \tag{3.3}$$

Diskretna Furijeova transformacija konačnog diskretnog signala $\{x(n)\}$ je definisana formulom [166]:

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{nk}, \;\; k = 0,..., N-1 \tag{3.4}$$

a inverzna diskretna Furijeova transformacija (eng. *Inverse Discrete Fourier Transform*, IDFT) formulom:

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) W_N^{-nk}, \;\; n = 0,..., N-1 \tag{3.5}$$

gde je $W_N = e^{-i2\pi/N}$ N-ti primitivni koren broja jedan, a N je dužina spektra, odnosno broj diskretnih frekvencija, i jednak je dužini signala.

Za generisanje spektra sa više elemenata, dovoljno je produžiti signal nultim elementima tj. dopuniti niz $\{x(n)\}$ na kraju sa nulama (eng. *zero padding*), čime se dobija spektar veće rezolucije sa manjim rastojanjima između tačaka na frekvencijskoj osi.

Za DFT realnih nizova važi osobina, slično kao i za Furijeovu transformaciju realnih nizova, da je $X(k)=X(N-k)^*$ (konjugovana simetrija), što ima za posledicu da je dovoljno izračunati prvih $N/2$ koeficijenata DFT-a.

3.3.3. Brza Furijeova transformacija (FFT)

Brza Furijeova transformacija (eng. *Fast Fourier Transform*, FFT) predstavlja naziv algoritama za efikasno računanje DFT-a. Vremenska kompleksnost DFT-a definisanog formulom (3.4) iznosi $O(N^2)$, dok je brzina FFT algoritma $O(N\log N)$ [167].

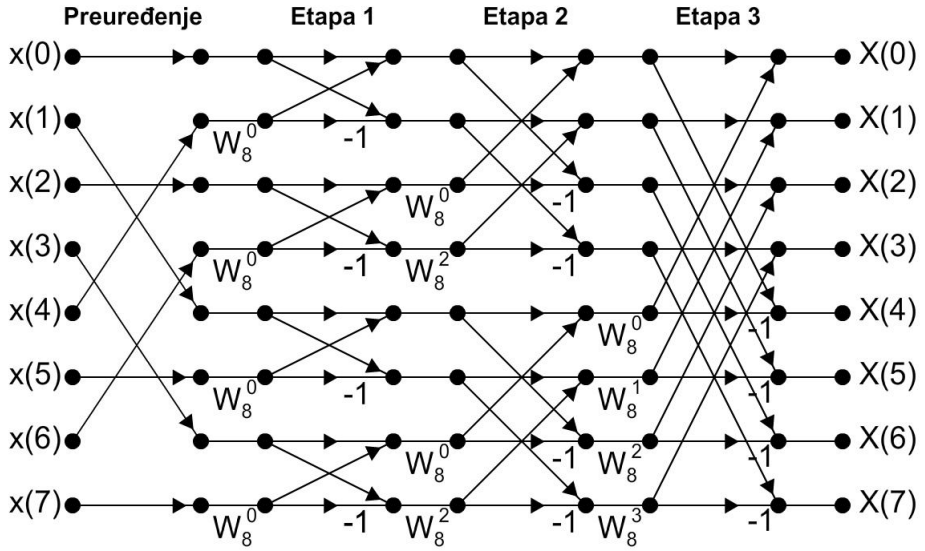
FFT algoritam sa razbijanjem po vremenu (eng. *FFT Decimation in Time*, FFT DIT) se zasniva na dekompoziciji izračunavanja DFT razbijanjem ulaznog niza $\{x(n)\}$ na podnizove (strategija „podeli pa vladaj“). Ako je dužina niza stepen broja dva, tj. $N=2^p$, onda se FFT naziva DIT algoritam sa osnovom dva (eng. *Decimation in time radix-2 FFT*) u kome vrši rekurzivna dekompozicija ulaznog niza $\{x(n)\}$ na dva dvostruko manja niza $\{x_1(n)\}$ parnih članova i $\{x_2(n)\}$ neparnih članova, dok se ne dođe do niza dužine dva. Elementi niza $\{X(k)\}$ dužine N su određeni rekurentnim formulama [168]:

$$X(k) = X_1(k) + W_N^k X_2(k), \;\; k = 0,..., N/2-1 \tag{3.6}$$

$$X(k+N/2) = X_1(k) - W_N^k X_2(k), \;\; k = 0,..., N/2-1 \qquad (3.7)$$

gde su $\{X_1(k)\}$ i $\{X_2(k)\}$ DFT nizovi dužina $N/2$ za ulazne nizove $\{x_1(n)\}$ i $\{x_2(n)\}$ respektivno, a $W_N = e^{-i2\pi/N}$ je N-ti primitivni koren broja jedan.

Ovakav FFT algoritam se, zbog svoje grafičke predstave, naziva *leptir* FFT DIT (eng. *FFT DIT butterfly*) [168]. Na slici 3.8 je prikazan dijagram toka *leptir* FFT algoritma na primeru niza od osam tačaka (N=8) sa tri etape.



Slika 3.8. Dijagram *leptir* FFT DIT algoritma za niz od 8 tačaka.

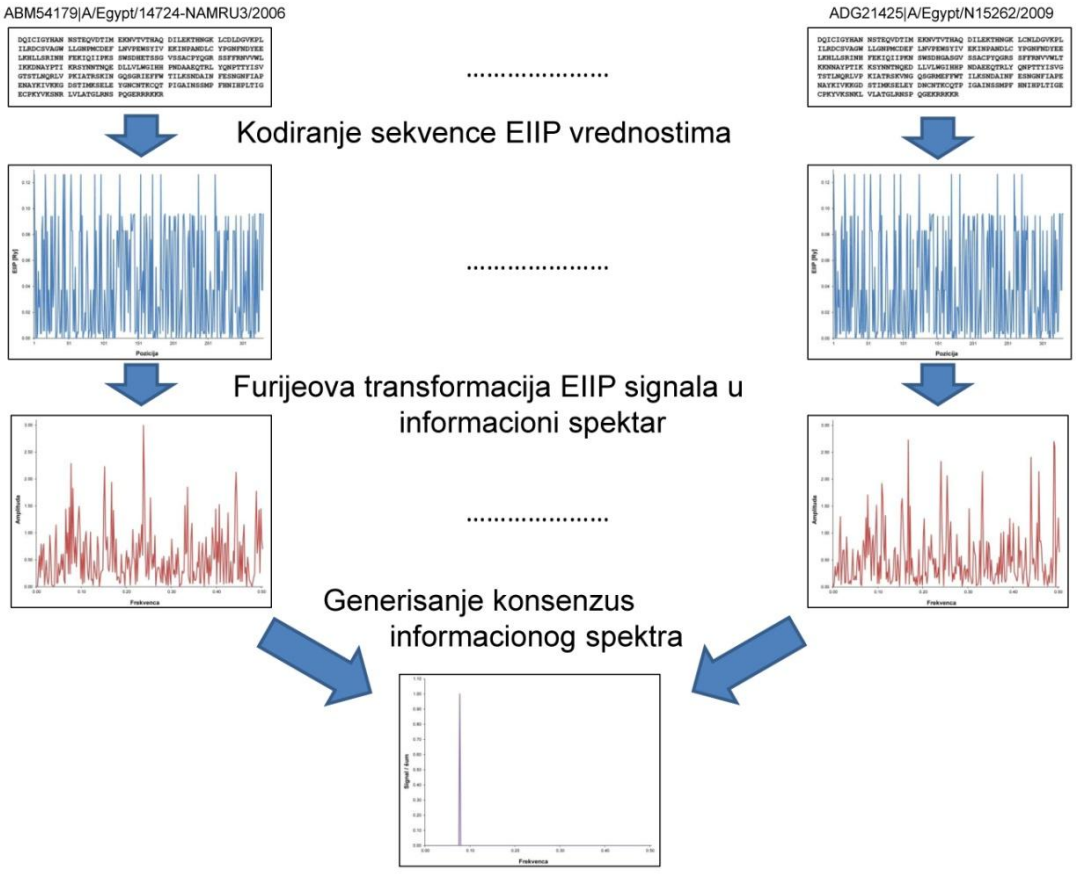
3.4. Metoda informacionih spektara (ISM)

Kako dalekodosežne interakcije između biomolekula predstavljaju bitan faktor koji utiče na biološke procese i predstavljaju suštinsku fizičko-hemijsku osobinu proteina i nukleotidnih sekvenci, one moraju biti uključene u analizi protein-protein i protein-DNK interakcija. EIIP predstavlja bitan fizički parametar koji određuje dalekodosežne osobine bioloških molekula i uzet je za osnovu metode informacionih spektara (eng. *informational spectrum method*, ***ISM***). Upoređivanje efikasnosti 226 strukturnih fizičko-hemijskih i termodinamičkih karakteristika aminokiselina u odnosu na informacije zapisane u primarnoj strukturi proteina, upotrebom ISM metode pokazano je da se jedino EIIP izdvaja kao univerzalan parametar za analizu odnosa između strukture i funkcije proteina, i međumolekulskih interakcija [169] .

ISM je metoda virtuelne spektroskopije za istraživanje protein-protein interakcija, protein-DNK interakcija, kao i za analizu odnosa strukture i funkcija kod proteina i nukleotidnih sekvenci. Metoda ISM se sastoji iz tri osnovna koraka:

1. Transformacija slovnog zapisa primarne strukture u niz brojeva, dodeljivanjem svakoj aminokiselini ili nukleotidu odgovarajuće EIIP vrednosti.
2. Transformacija dobijenog niza brojeva u informacioni spektar primenom Furijeove transformacije.
3. Konsenzus-spektralna analiza koja omogućava identifikaciju karakterističnih frekventnih komponenti informacionih spektara molekula koje su značajne za određenu biološku funkciju ili interakciju sa drugim molekulima.

Na slici 3.9. je shematski predstavljen algoritam ISM metode, sastavljen od osnovnih koraka, od kodiranja pojedinačne sekvence u EIIP signal i transformacije u informacioni spektar, do generisanje konsenzusa svih spektara.

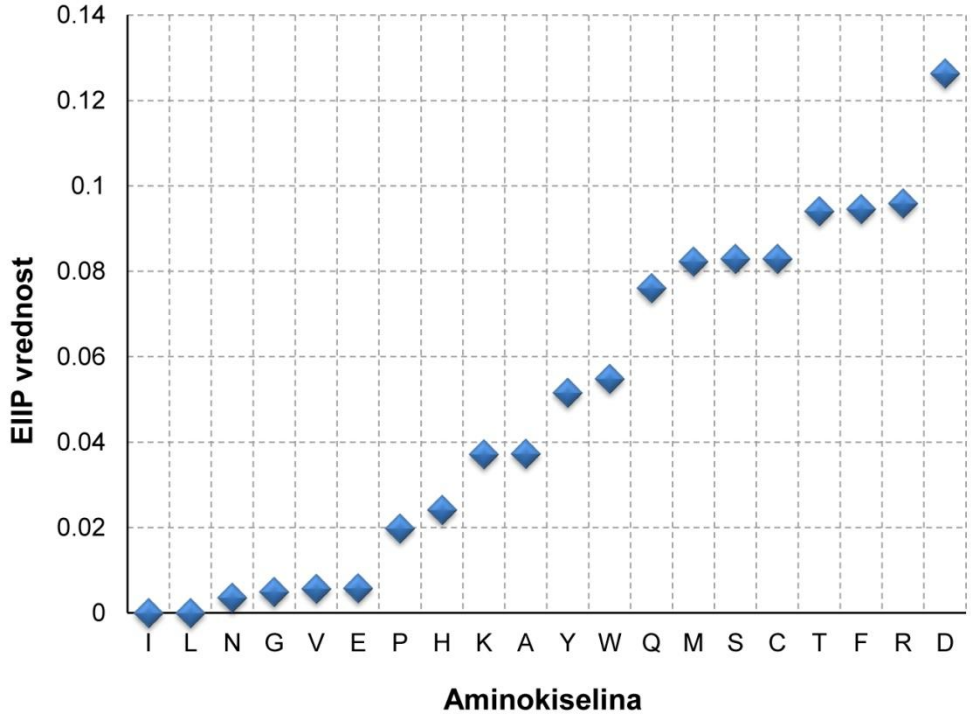


Slika 3.9. Shematska prikaz algoritma ISM metode.

ISM metodom se prvo kodira zapis sekvence u signal (numerički niz) tako što svakoj aminokiselini iz primarnog zapisa sekvence dodeli odgovarajuća EIIP vrednost (tabela 3.6, slika 3.10).

Tabela 3.6. EIIP vrednosti aminokiselina.

Kiselinski ostatak	**Troslovni zapis**	**Jednoslovni zapis**	**EIIP vrednost**
Alanin	Ala	A	0.03731
Arginin	Arg	R	0.09593
Asparagin	Asn	N	0.00359
Asparaginska kiselina	Asp	D	0.12630
Cistein	Cys	C	0.08292
Glutaminska kiselina	Glu	E	0.00580
Glutamin	Gln	Q	0.07606
Glicin	Gly	G	0.00499
Histidin	His	H	0.02415
Izoleucin	Ile	I	0.00000
Leucin	Leu	L	0.00000
Lizin	Lys	K	0.03718
Metionin	Met	M	0.08226
Fenilalanin	Phe	F	0.09460
Prolin	Pro	P	0.01979
Serin	Ser	S	0.08292
Treonin	Thr	T	0.09408
Triptofan	Trp	W	0.05481
Tirozin	Tyr	Y	0.05159
Valin	Val	V	0.00569



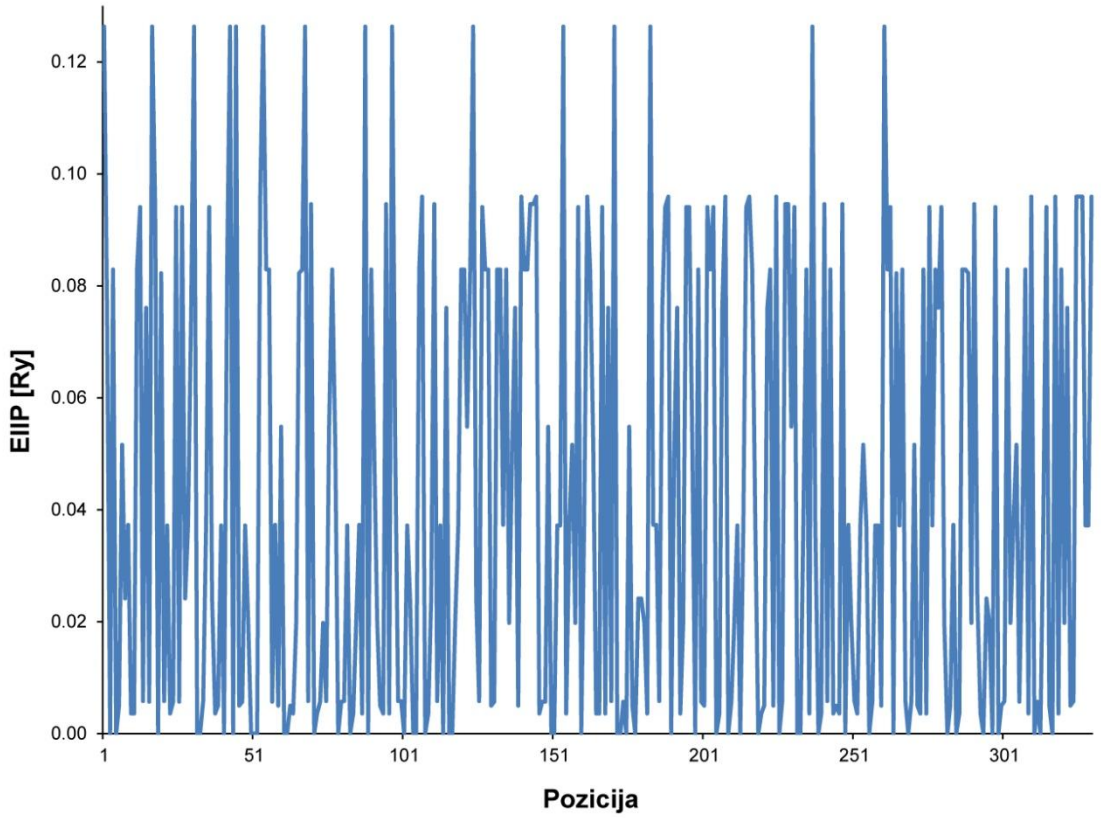
Slika 3.10. Grafička prezentacija EIIP vrednosti za aminokiseline.

Tabela 3.7. EIIP vrednosti nukleinskih kiselina.

Nukleotidna baza	**Jednoslovni zapis**	**EIIP vrednost**
Adenin	A	0.1260
Citozin	C	0.1340
Guanin	G	0.0806
Timin	T	0.1335
Uracil	U	0.0550

Primer zapisa primarne strukture proteina HA1 H5N1 (GenBank: ABM54179|A/Egypt/14724-NAMRU3/2006):

```
DQICIGYHAN NSTEQVDTIM EKNVTVTHAQ DILEKTHNGK LCDLDGVKPL ILRDCSVAGW
LLGNPMCDEF LNVPEWSYIV EKINPANDLC YPGNFNDYEE LKHLLSRINH FEKIQIIPKS
SWSDHETSSG VSSACPYQGR SSFFRNVVWL IKKDNAYPTI KRSYNNTNQE DLLVLWGIHH
PNDAAEQTRL YQNPTTYISV GTSTLNQRLV PKIATRSKIN GQSGRIEFFW TILKSNDAIN
FESNGNFIAP ENAYKIVKKG DSTIMKSELE YGNCNTKCQT PIGAINSSMP FHNIHPLTIG
ECPKYVKSNR LVLATGLRNS PQGERRRKKR
```



Slika 3.11. Primer EIIP signala proteina HA1 H5N1 influence A virusa (GenBank: ABM54179|A/Egypt/14724-NAMRU3/2006).

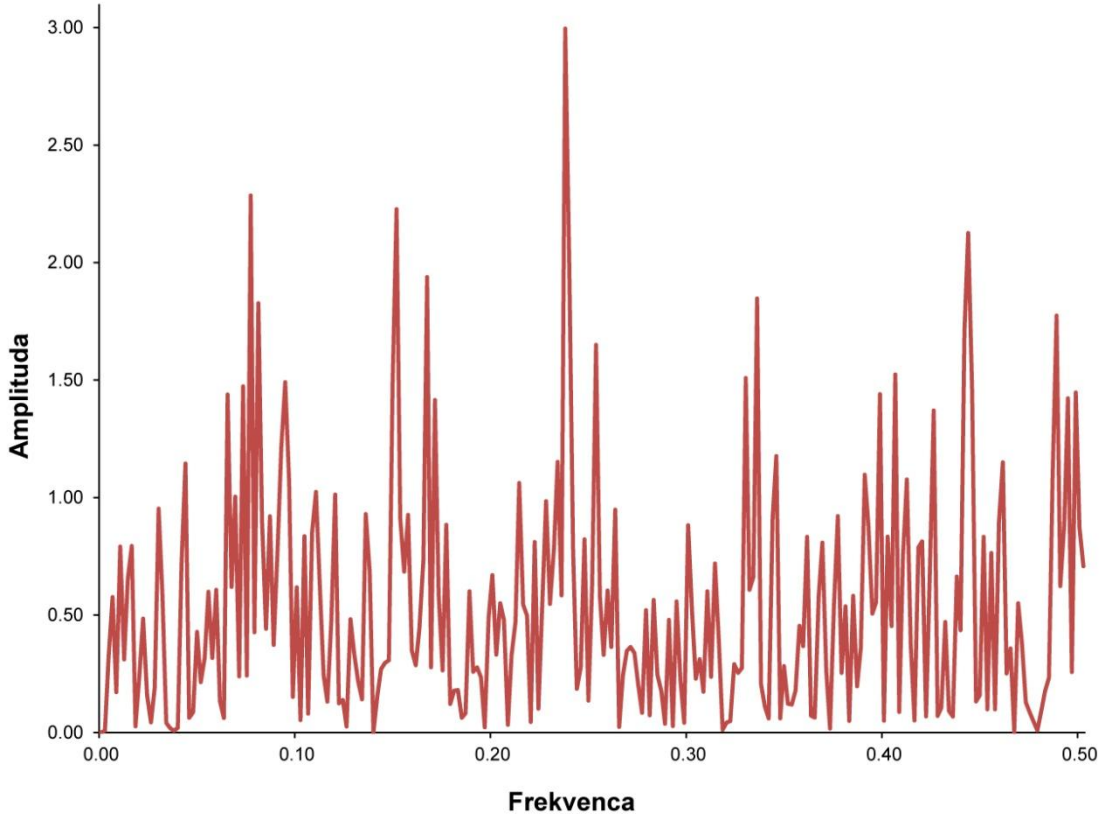
Zatim se signal transformiše u spektar diskretnom Furijeovom transformacijom:

$$X(n) = \sum_{m=1}^{N} x(m) e^{-i2\pi n(m-1)/N}, \quad n = 1,..., N/2 \tag{3.8}$$

Gde je *x(m)* *m*-ti član EIIP signala, *X(n)* je *n*-ti koeficijent diskretne Furijeove transformacije, *N* je dužina signala odnosno ukupan broj elemenata u EIIP numeričkom nizu. Ovi koeficijenti opisuju amplitudu, fazu i frekvencu sinusoida koje sačinjavaju originalni signal. Potpuna informacija o originalnoj sekvenci je sadržana u amplitudnom i faznom spektru.

U slučaju analize proteina, relevantna informacija je predstavljena u spektru gustine energije [170] koji je definisan sledećom formulom:

$$S(n) = X(n)X^{*}(n) = \left|X(n)\right|^{2}, \quad n = 1,..., N/2 \tag{3.9}$$

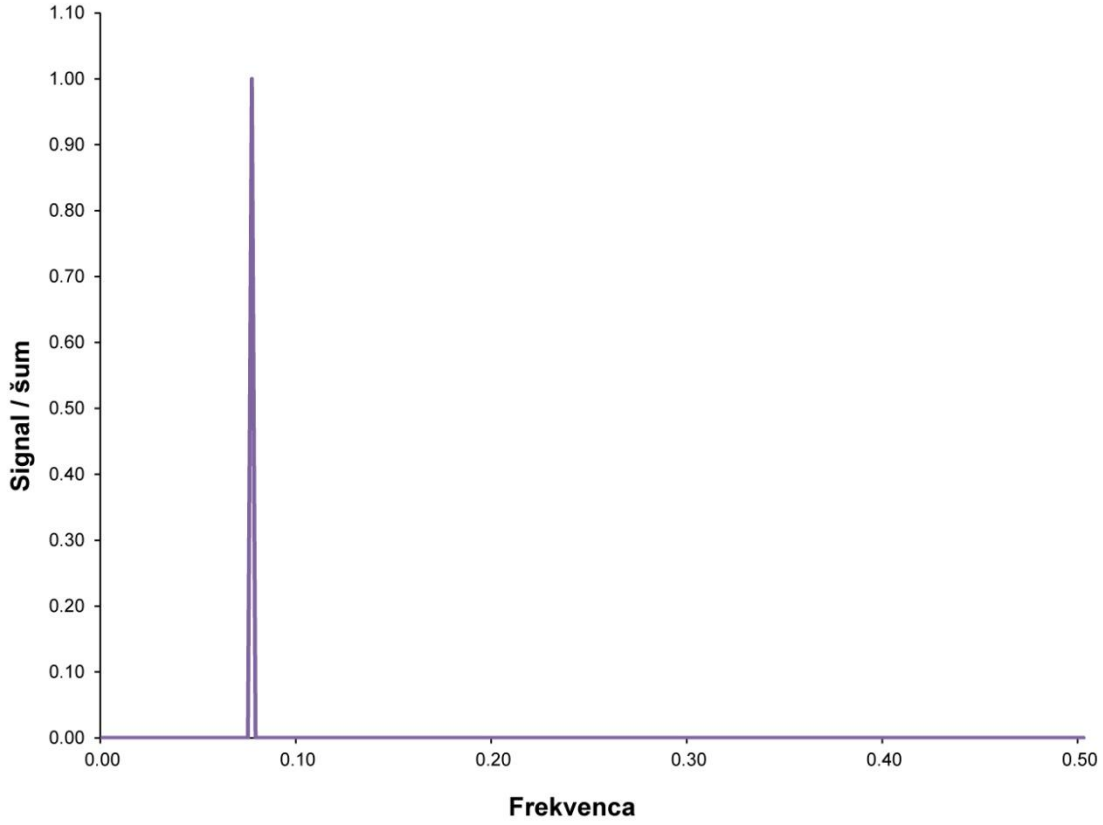


Slika 3.12. Primer informacionog spektra proteina HA1 H5N1 influence A virusa (GenBank: ABM54179|A/Egypt/14724-NAMRU3/2006).

Na ovaj način se sekvenca može analizirati kao diskretan signal. Uz pretpostavku da su tačke signala ekvidistantne sa ukupnom distancom d=1, uz osobinu da je spektar signala realnih brojeva (EIIP vrednosti) simetričan, maksimalna frekvenca u spektru definisanim formulom (3.9) je $F = 1/2\ d = 0.5$. Raspon frekvenci je na taj način nezavisan od dužine sekvenci (broja aminokiselina u sekvenci). Dužina sekvence utiče samo na rezoluciju spektra. Ukupna rezolucija spektra dužine N je $f = 1/N$, gde n-ti element spektra odgovara frekvenci $f(n) = n\ f = n/N$. Na taj način početna informacija definisana sekvencom aminokiselina može biti predstavljena u formi informacionog spektra (**IS**), koji je predstavljen nizom frekvenci i njihovih amplituda.

Frekvence informacionog spektra odgovaraju raspodeli strukturnih motiva definisanih fizičkohemijskim osobinama koje određuju biološku funkciju proteina. Pri upoređivanju proteina koji dele istu biološku ili biohemijsku funkciju, ISM metoda omogućava detekciju parova kôd-frekvenca koji su specifični za njihove zajedničke biološke osobine, ili odgovaraju njihovoj specifičnoj interakciji. Ova zajednička

informaciona karakteristika sekvenci je definisana kros-spektrom (eng. *Cross-Spectrum*, ***CS***) tj. konsenzus informacionim spektrom (eng. *Consensus Informational Spectrum*, ***CIS***).



Slika 3.13. Primer konsenzus informacionog spektra svih 526 objavljenih sekvenci HA1 H5N1 influence A virusa, u NCBI bazi, izolovanih u Egiptu između 2006 i 2011. godine.

Konsenzus informacioni spektar za *M* spektara je definisan formulom:

$$C(j) = \prod_{i=1}^{M} S(i, j), \quad j = 1,\ldots, N/2 \tag{3.10}$$

gde je $S(i, j)$ j-ti element i-tog spektra, $C(j)$ je j-ti element kros-spektra. Tako definisan CIS je Furijeova transformacija konvolucije spektara. Na taj način se eliminiše svaka frekvenca (spektralna komponenta) koja nije prisutna u svim informacionim spektrima koji se upoređuju. *Pikovi* u kros-spektru su zajedničke frekventne komponente analiziranih sekvenci (pikovi su vrhovi tj. lokalni maksimumi spektra). Mera sličnosti sekvenci određenog pika u kros-spektru je vrednost odnosa signala i šuma (eng. *signal*

to noise ratio, S/N) na frekvenci tog pika i predstavlja odnos između jačine signala na određenoj frekvenci informacionog spektra i ukupne vrednosti na celom spektru.

Kada se izračuna kros-spektar za grupu proteina koji imaju različite primarne strukture, i detektuju se striktno definisani spektralni pikovi, to znači da analizirani proteini učestvuju u međusobnoj interakciji ili imaju zajedničku biološku funkciju.

ISM metoda je uspešno primenjen u analizi odnosa struktura i funkcija raznih proteina i DNK sekvenci, kao i u *de novo* dizajnu biološko aktivnih peptida [171-191]. ISM predstavlja jednostavan i efikasan alat za: (i) analizu odnosa struktura i funkcija proteina i identifikaciju funkcionalno značajnih domena koji predstavljaju kandidate za terapeutsku metu, (ii) kreiranje *in silico* interaktoma i pretraživanje mogućih interaktora određenih proteina i nukleotidnih sekvenci u bazama sekvenci, (iii) procenu bioloških efekata mutacija.

3.5. Algoritmi hijerarhijskog klasterisanja u filogenetskoj analizi

U filogenetskoj analizi, metode za konstrukciju filogenetskih stabala zasnovane na rastojanjima su veoma brze i primenjuju se na grupi sekvenci koje dele prepoznatljivu sličnost. Metode za generisanje filogenetskog stabla zasnovane na rastojanju se sastoje iz dva osnovna koraka: u prvom se generiše matrica rastojanja između svake dve sekvence iz skupa koji se analizira, u drugom koraku se transformiše matrica rastojanja u stablo koristeći neki od aglomerativnih metoda hijerarhijskog klasterisanja. Najčešće korišćene metode su UPGMA [26] i NJ [28].

Aglomerativne metode hijerarhijskog klasterisanja grade klastere „odozdo na gore“, odnosno iterativno spajaju najbliže skupove, gde je na početku svaki elemenat jedan skup, dok se svi elementi ne spoje u jedan skup.

Osnovni koraci aglomerativno hijerarhijskih algoritama su:

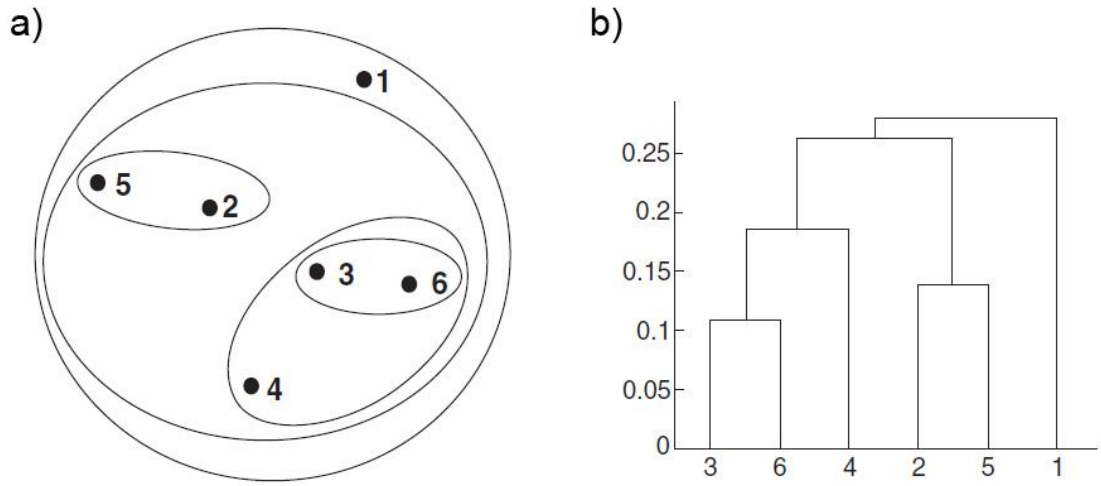
Ulaz: skup elemenata

Izlaz: dendogram

1. Izračunati matricu rastojanja između elemenata
2. Inicijalizacija grupa: svaki element je jedna grupa
3. Dok postoji više od jedne grupe ponavljaj:
    a. Spojiti dve najbliže grupe u novu grupu

b. Preračunati rastojanja nove grupe sa svima ostalima

c. Izbaciti spojene grupe.



Slika 3.14. Primer hijerarhijskog klasterisanja šest tačaka. a) klasteri, b) dendogram.

Hijerarhijski algoritmi konstruišu skup ugnježdenih grupa elemenata koji se predstavlja dendogramom, gde dužina grana između čvorova odgovara rastojanjima između podskupova. Osnovna osobina ovih algoritama je da definišu više klasterisanja istog skupa elemenata. Na svakom nivou, odnosno koraku algoritma, kreirani skupovi su disjunktni i čine jedno klasterisanje početnih elemenata.

Vremenska komleksnost aglomerativnih algoritama je O(N^2log$N$), a prostorna je O(N^2), gde je N broj elemenata.

Postoje različite tehnike izračunavanja rastojanja između nove kreirane i ostalih grupa elemenata.

3.5.1. UPGMA metoda

UPGMA metoda (eng. *Unweighted Pair Group Method with Arithmetic Mean*) [26] je aglomerativni hijerarhijski algoritam za klasterisanje pomoću kojeg se rastojanje između dva skupa računa kao srednja vrednost rastojanja između svaka dva elementa prvog i drugog skupa. Rastojanje između skupova C_i i C_j je:

$$d_{i,j} = d(C_i, C_j) = \frac{\sum_{x \in C_i, y \in C_j} d(x, y)}{m_i \, m_j} \tag{3.11}$$

gde je m_i veličina skupa C_i, m_j je veličina skupa C_j.

Ako je $C_k=C_i \cup C_j$, a C_l bilo koji skup, onda je na osnovu formule (3.11) rastojanje između skupova C_k i C_l zadato formulom:

$$d_{k,l} = \frac{d_{i,l} m_i + d_{j,l} m_j}{m_i + m_j} \tag{3.12}$$

Algoritam UPGMA:

Ulaz: skup sekvenci

Izlaz: korensko stablo

1. Za svaku sekvencu *i*, kreiraj njen skup *Ci*
2. Za svaku sekvencu postavi list stabla na poziciju $y=0$
3. Dok postoji više od dva aktivna skupa:
    a. Pronađi dva najbliža skupa *i* i *j*, odnosno da $d_{i,j}$ bude minimalno
    b. Definiši novi skup *k*: $C_k=C_i \cup C_j$ i izračunaj rastojanja $d_{k,l}$ za svako *l* prema formuli (3.12)
    c. Kreiraj čvor *k* sa čvorovima decom *i* i *j*, zatim ga postavi na poziciju $y=d_{i,j}/2$
    d. Dodaj *k* u aktivne skupove, obriši *i* i *j*.
4. Za dva poslednja skupa *i* i *j*, postavi koren stabla na poziciju $y=d_{i,j}/2$.

Algoritam UPGMA kreira korensko stablo u kojem su svi listovi na istoj visini. U takvom stablu se dužine grana mogu posmatrati kao vremena izmerena molekulskim satom sa konstantnom brzinom. Pretpostavlja se da je divergencija sekvenci u UPGMA stablu odvija konstantom brzinom u svakoj tački. To je ekvivalentno tvrđenju da je iz istog čvora zbir dužina grana do bilo kog lista podjednak.

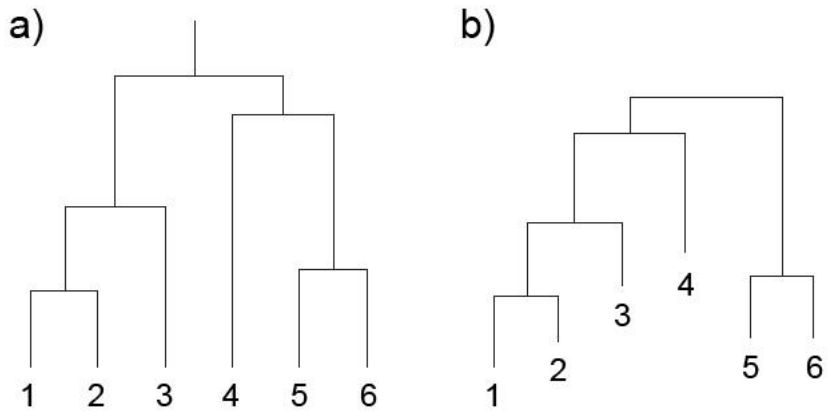
Metrika *d* je **ultrametrika** ako pored osnovnih osobina (nenegativnost, ekvivalentnost jednakosti, simetričnost, nejednakost trougla) zadovoljava dodatni uslov:

$d(x,y) \leq \max(d(x,z),d(y,z))$, za svake tri tačke x,y,z

koji je ekvivalentan uslovu „uslov tri tačke“: tačke x,y,z se mogu preimenovati tako da važi: $d(x,y) \leq d(x,z) = d(y,z)$, tj. dva od tri rastojanja su međusobno jednaka i veća ili jednaka od trećeg.

Ultrametrika i UPGMA su povezani sledećim tvrđenjem: Ako je rastojanje između sekvenci ultrametrika, onda stablo konstruisano UPGMA algoritmom ispravno

rekonstruiše originalnu stablo zadato matricom rastojanja, odnosno vertikalna rastojanja listova (preko zajedničkog čvora) u stablu odgovaraju rastojanjima iz matrice rastojanja [192].



Slika 3.15. Primer stabla konstruisanog: a) UPGMA algoritmom, b) NJ algoritmom.

3.5.2. Metoda spajanja suseda (NJ)

U slučaju kada nije ispunjen uslov ultrametrike, UPGMA neće ispravno rekonstruisati stablo. U realnosti molekulski sat ima različite brzine za različite vrste organizama. Zato je uslov ultrametrike oslabljen i definisan uslov aditivnosti.

Metrika *d* je **aditivna** ako pored osnovnih osobina metrike zadovoljava dodatni uslov aditivnosti:

$d(x,y)+d(u,v) \leq \max(d(x,u)+d(y,v), d(x,v)+d(y,u))$, za svake četiri tačke x,y,u,v koji je ekvivalentan „uslovu četiri tačke“. Naime, tačke x,y,u,v se mogu preimenovati tako da važi: $d(x,y)+d(u,v) \leq d(x,u)+d(y,v) = d(x,v)+d(y,u)$, tj. dva od tri zbira rastojanja su međusobno jednaka i veća ili jednaka od trećeg. Primer aditivne mere je L_1 metrika Minkovskog [193].

NJ metoda (eng. *Neighbor-Joining*) [28] je aglomerativni hijerarhijski algoritam za klasterisanje koji ispravno rekonstruiše stablo sa aditivnim rastojanjima, tj. stablo gde je rastojanje između svaka dva lista jednako zbiru dužina grana koje ih spajaju [37].

NJ funkcioniše tako što pronalazi parove susednih listova (koji ne moraju biti i najbliži), odnosno onih koji imaju zajedničkog roditelja i rekonstruiše grane stabla. Rastojanje između kreiranog roditelja čvora *k* sa decom *i*, *j* i bilo kog čvora *l* je:

$$d_{k,l} = (d_{i,l} + d_{j,l} - d_{i,j})/2 \quad (3.13)$$

Dva najbliža čvora u stablu ne moraju biti i susedna. Da bi se rešio ovaj problem i pronašli susedi, Satou i Nei su uveli novo rastojanje koje kompenzuje dužine dugačkih grana [28]:

$$D_{i,j} = d_{i,j} - (r_i + r_j) \quad (3.14)$$

gde je: $r_i = \frac{1}{|L|-2}\sum_{k \in L} d_{i,k}$

Algoritam NJ:

Ulaz: skup sekvenci

Izlaz: nekorensko stablo

1. Za svaku sekvencu, kreiraj njen list *i* i stavi ga u aktivan skup čvorova
2. Dok postoji više od dva aktivna lista:
    a. Pronađi dva najbliža aktivna lista *i* i *j* tj. da $D_{i,j}$ definisano formulom (3.14) bude minimalno
    b. Definiši novi čvor *k* i izračunaj rastojanja $d_{k,l}$ za svako aktivno *l* prema formuli (3.13)
    c. Kreiraj čvor *k* sa granama do *i* i *j*, sa dužinama $d_{i,k}=(d_{i,j}+r_i-r_j)/2$, $d_{j,k}= d_{i,j}-d_{i,k}$
    d. Dodaj *k* u aktivne čvorove, obriši *i* i *j*.
3. Za dva poslednja čvora *i* i *j*, dodaj granu između *i* i *j* dužine $d_{i,j}$.

4. Rezultati

4.1. Novi algoritam za filogenetsku analizu proteina (ISTREE)

4.1.1. Informaciono filogenetsko stablo i nova mera proteinskih rastojanja zasnovana na ISM metodi

Bitna mana filogenetske analize zasnovane na MSA je da sličnost sekvenci ne podrazumeva automatski i sličnost bioloških funkcija. Na primer, dve proteinske sekvence koje se razlikuju u jednoj mutaciji, koja je letalna za biološku funkciju, biće filogenetski bliske. Sa druge strane, dva proteina koja se razlikuju u više mutacija, gde ta kombinacija mutacija ne utiče na promenu biološke funkcije, biće filogenetski udaljeni. Da bi se ove mane prevazišle i da bi se unapredila funkcionalna analiza sekvenci, razvijene su nove mere proteinskih rastojanja, zasnovane na metodi ISM, nezavisne od MSA.

Novi princip filogenetske analize zasnovan na informacionom spektru, nazvan je ISTREE (eng. *Informational Spectrum-based Phylogenetic Tree*).

U zavisnosti od funkcije koja se analizira i informacije koja se izvodi, definisana su tri tipa rastojanja.

4.1.1.1. Rastojanje na pojedinačnoj frekvenci

Kada je pronađena karakteristična frekvenca F u konsenzus informacionom spektru grupe proteina, gde F reprezentuje određenu biološku karakteristiku i ako je cilj analize izvođenje stabla koje prikazuje filogenezu pojedinačne informacije, onda se rastojanje između sekvenci može definisati kao apsolutna razlika amplituda na frekvenci F informacionih spektara sekvenci.

Neka su X_1 i X_2 dve sekvence, S_1 i S_2 njihovi odgovarajući informacioni spektri. Neka je F karakteristična frekvenca, a $A_1(F)$ i $A_2(F)$ amplitude na frekvenci F spektara S_1 i S_2. Rastojanje između sekvenci X_1 i X_2 je definisano na sledeći način:

$$d_1(X_1, X_2) = \left| A_1(F) - A_2(F) \right| \tag{4.1}$$

Ako je P skup svih vrednosti amplituda $A(F)$ za svaku sekvencu X, kako je rastojanje d_1 Euklidsko rastojanje na skupu realnih brojeva R, i $P \subseteq R$, onda je d_1 validna metrika na skupu P, odnosno zadovoljava sledeće osobine:

1. $d_1(x,y) \geq 0$ (nenegativnost)
2. $d_1(x,y) = 0 \Leftrightarrow x=y$ (ekvivalentnost jednakosti)
3. $d_1(x,y) = d_1(y,x)$ (simetrija)
4. $d_1(x,z) \leq d_1(x,y) + d_1(y,z)$ (nejednakost trougla)

Pored toga, rastojanje d_1 je validna aditivna evolucijska mera na skupu P jer zadovoljava aditivnost, tj. uslov četiri tačke: dve od tri sume: $d_1(x,y)+d_1(z,w)$, $d_1(x,z)+d_1(y,w)$, $d_1(x,w)+d_1(y,z)$, su jednake i veće od treće sume. Kako NJ metoda garantovano nalazi stablo koje tačno odgovora ulaznoj matrici rastojanja, u slučaju kada rastojanje zadovoljava uslov aditivnosti [37], NJ algoritam će ispravno rekonstruisati stablo sa rastojanjem d_1.

4.1.1.2. Rastojanje odnosa amplituda na dve frekvence

Ako su pronađene dve karakteristične frekvence F_1 i F_2 u konsenzus informacionom spektru određene grupe proteina, gde F_1 i F_2 predstavljaju određene biološke karakteristike i ako je cilj dobijanje informacije koja odgovara meri prelaza između ove dve karakteristike, onda se može definisati rastojanje između sekvenci kao apsolutna razlika odnosa amplituda.

Neka su X_1 i X_2 dve sekvence, S_1 i S_2 njihovi odgovarajući informacioni spektri. Neka su F_1 i F_2 dve karakteristične frekvence, $A_1(F_1)$ i $A_2(F_1)$ amplitude na frekvenci F_1 u spektrima S_1 i S_2, slično $A_1(F_2)$ i $A_2(F_2)$ amplitude na frekvenci F_2. Rastojanje između sekvenci X_1 i X_2 se može definisati na sledeći način:

$$d_2(X_1, X_2) = \left| \frac{A_1(F_1)}{A_1(F_2)} - \frac{A_2(F_1)}{A_2(F_2)} \right| \tag{4.2}$$

Ako je P skup svih vrednosti odnosa amplituda $A(F_1)/A(F_2)$ za svaku sekvencu X, slično kao i za d_1, rastojanje d_2 je validna metrika na skupu P i zadovoljava aditivnost, što za posledicu ima da će NJ algoritam ispravno rekonstruisati stablo sa d_2 rastojanjem.

4.1.1.3. Rastojanje na celom spektru

Ako je cilj analize konstrukcija stabla koje će uzeti u obzir punu informaciju iz konsenzus spektra određene grupe proteina, onda se mogu uzeti u obzir celi spektri i rastojanje se može definisati kao rastojanje Minkovskog L_1, (Menhetn metrika). Neka su X_1 i X_2 dve sekvence, $S_1=\{S_1(n)\}$ i $S_2=\{S_1(n)\}$, n = 1, 2, ..., N/2, njihovi odgovarajući informacioni spektri i N rezolucija spektara. Rastojanje između X_1 i X_2 se može definisati na sledeći način:

$$d_3(X_1, X_2) = \frac{1}{N}\sum_{n=1}^{N/2}\left|S_1(n) - S_2(n)\right| \qquad (4.3)$$

Rastojanje d_3 definisano formulom (4.3) je L_1 metrika na vektorskom prostoru $R^{N/2}$ informacionih spektara, pomnožena sa konstantom 1/N, gde su svi spektri iste rezolucije. Kao posledica nejednakosti Minkovskog [193], d_3 je validna metrika.

4.1.2. Osobine novog rastojanja

(i) ISM filogenetski pristup nije zasnovan na višestrukom poravnavanju (MSA) i ne koristi nijedan supstitucioni model.

(ii) ISM rastojanje je osetljivo na poziciju mutacije i tip supstitucionog ostatka, za razliku od standardnih pristupa gde se sve pozicije posmatraju podjednako, kao što je u Dayhoff [194] i Jones-Taylor-Thornton (JTT) [195] supstitucionim modelima, gde su proteinska rastojanja osetljiva samo na tip mutacije. Supstitucija istom aminokiselinom na različitim pozicijama u mutiranim sekvencama uzrokuje promenu EIIP vrednosti na tim istim različitim pozicijama u vektoru signala $\{x(m)\}$, što dalje uzrokuje različite promene na celim informacionim spektrima $\{S(n)\}$ mutiranih sekvenci definisanih formulama (3.8) i (3.9). To znači da je ISM rastojanje, definisano formulama (4.1), (4.2) i (4.3) između mutirane i nemutirane sekvence, direktno zavisi od pozicije pojedinačne mutacije.

Na slici 4.1.1 i u tabeli 4.1.1 date su srednje vrednosti i standardna odstupanja ISM rastojanja na celom spektru d_3 i JTT rastojanja, između nemutiranih i mutiranih sekvenci. Rastojanja su predstavljena kao funkcije supstitucija po svih 110 pozicija u sekvenci humanog insulina, aminokiselinama I, A i D. Izabrane kiseline I, A i D

reprezentuju minimalnu, srednju, i maksimalnu EIIP vrednost. ISM rastojanja pokazuju veću varijaciju i osetljivost na pozicije u odnosu na JTT rastojanja.

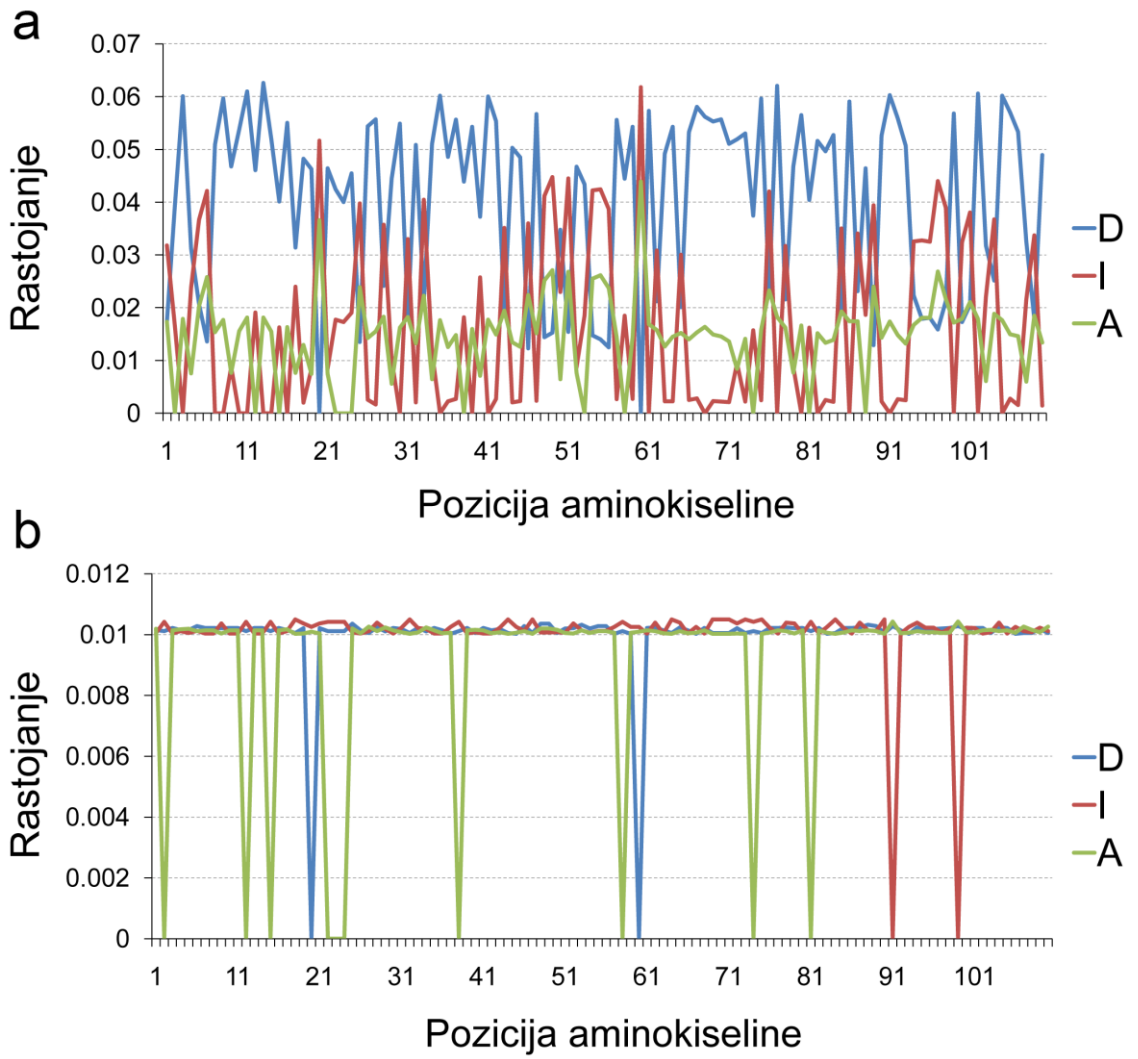
Srednje vrednosti i standardna odstupanja ISM rastojanja d_2 odnosa amplituda na frekvencama F(0.236) i F(0.076), između HA1 proteina iz H5N1 HPAIV (GenBank: ABW37431) i odgovarajućih mutiranih sekvenci date su u tabeli 4.1.2. Mutirane sekvence su dobijene uvođenjem pojedinačne supstitucije redom na svaku od 32 nekonzervirane pozicije proteina HA1, aminokiselinama I, P, K, Y, Q, T i D. Rezultati pokazuju da pozicije utiču na osetljivost na supstitucije različitim aminokiselinama (slika 4.1.2), i da se uticaj svake aminokiseline razlikuje u zavisnosti od pozicije (tabela 4.1.2).

Za svaki skup od 32 sekvence dobijene supstitucijom pojedinačne aminokiseline (I, P, K, Y, Q, T i D) na svakoj od 32 nekonzervirane pozicije HA1 proteina, generisano je filogenetsko stablo ISM algoritmom i standardnim pristupom. Rezultat na slikama 4.1.3a-g pokazuje da na strukturu ISM stabla značajno utiče pojedinačna mutacija, za razliku od standardnog stabla zasnovanog na MSA koje nije osetljivo na pojedinačnu aminokiselinsku supstituciju (slike 4.1.3.h-n).

(iii) ISM filogenetska analiza je osetljiva na pojedinačnu mutaciju i poziciju mutacije. Generisan je skup od 110 mutiranih sekvenci tako što je uvedena supstitucija aminokiselinom I na svaku pojedinačnu poziciju u sekvenci humanog insulina. Istom procedurom, sa kiselinama A i D, su generisane još dve grupe sekvenci. Konstruisana su filogenetska stabla za ove grupe sekvenci primenom standardnog i ISM pristupa korišćenjem rastojanja na celom spektru d_3. Stabla generisana ISM pristupom (slike 4.1.4c,f,i) prikazuju veću raznovrsnost i osetljivost na pojedinačnu poziciju mutacije, za razliku od stabla dobijena standardnim pristupom (slike 4.1.4a,b,d,e,g,h).

(iv) Izračunavanje EIIP vrednosti za organske molekule, uključujući aminokiseline, je zasnovano isključivo na bruto hemijskoj formuli. To za posledicu ima da su EIIP vrednosti dva izomera nekog organskog molekula jednake. To znači da aminokiseline leucin i izoleucin imaju iste EIIP vrednosti koje su jednake 0 Ry. Prema ISM konceptu, izmena ostatka u nekom proteinu aminokiselinom sa istom EIIP vrednošću neće uticati na promenu IS osobina koje reprezentuju dalekodosežne

karakteristike proteina. Drugim rečima, takve mutacije ne menjaju poziciju proteina u ISM filogenetskom stablu.



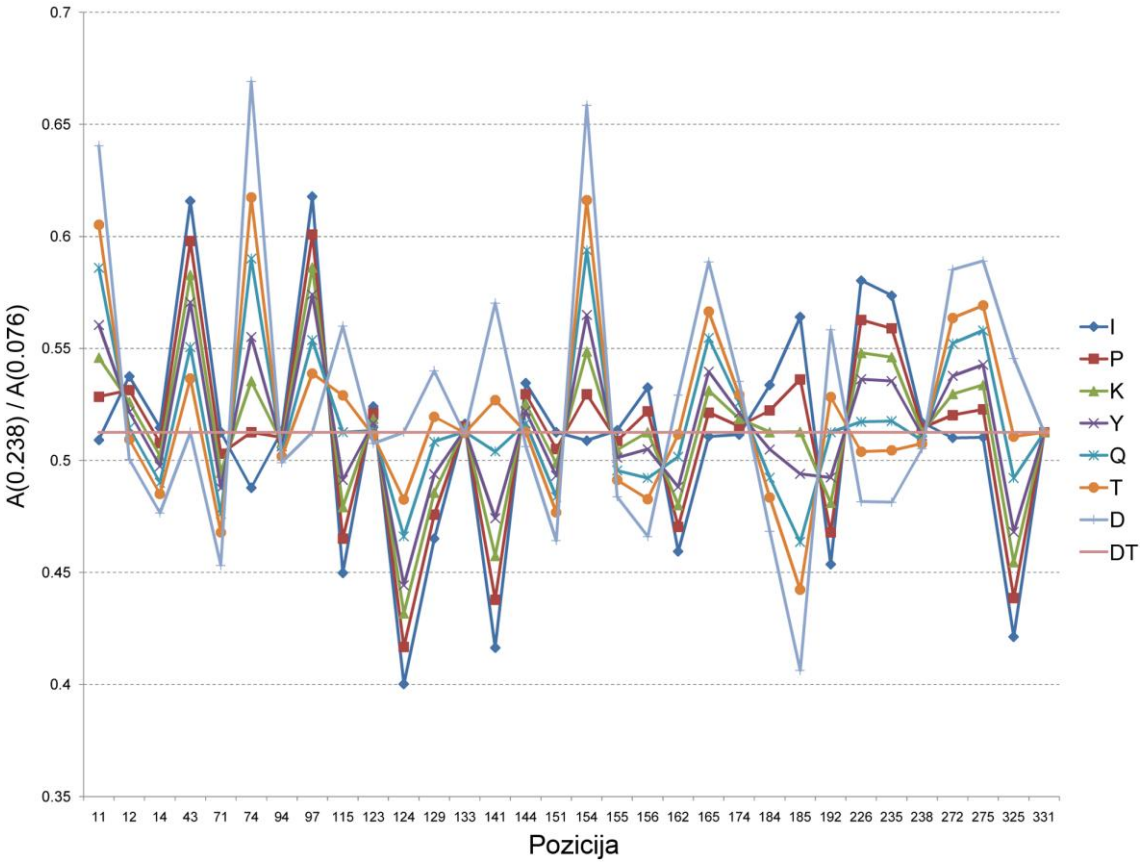
Slika 4.1.1. Grafička prezentacija varijacije a) ISM rastojanja na celom spektru i b**)** JTT rastojanja između proteinske sekvence humanog insulina i mutiranih sekvenci kao funkcija pojedinačnih mutacija po svih 110 aminokiselinskih pozicija humanog insulina (GenBank: AAA59172,1) sa kiselinama I, A i D.

Tabela 4.1.1. Srednje vrednosti i standardna odstupanja ISM i JTT rastojanja između humanog insulina i odgovarajućih mutiranih sekvenci uvođenjem pojedinačne supstitucije na svaku od 110 pozicija insulina, aminokiselinama I, A i D.

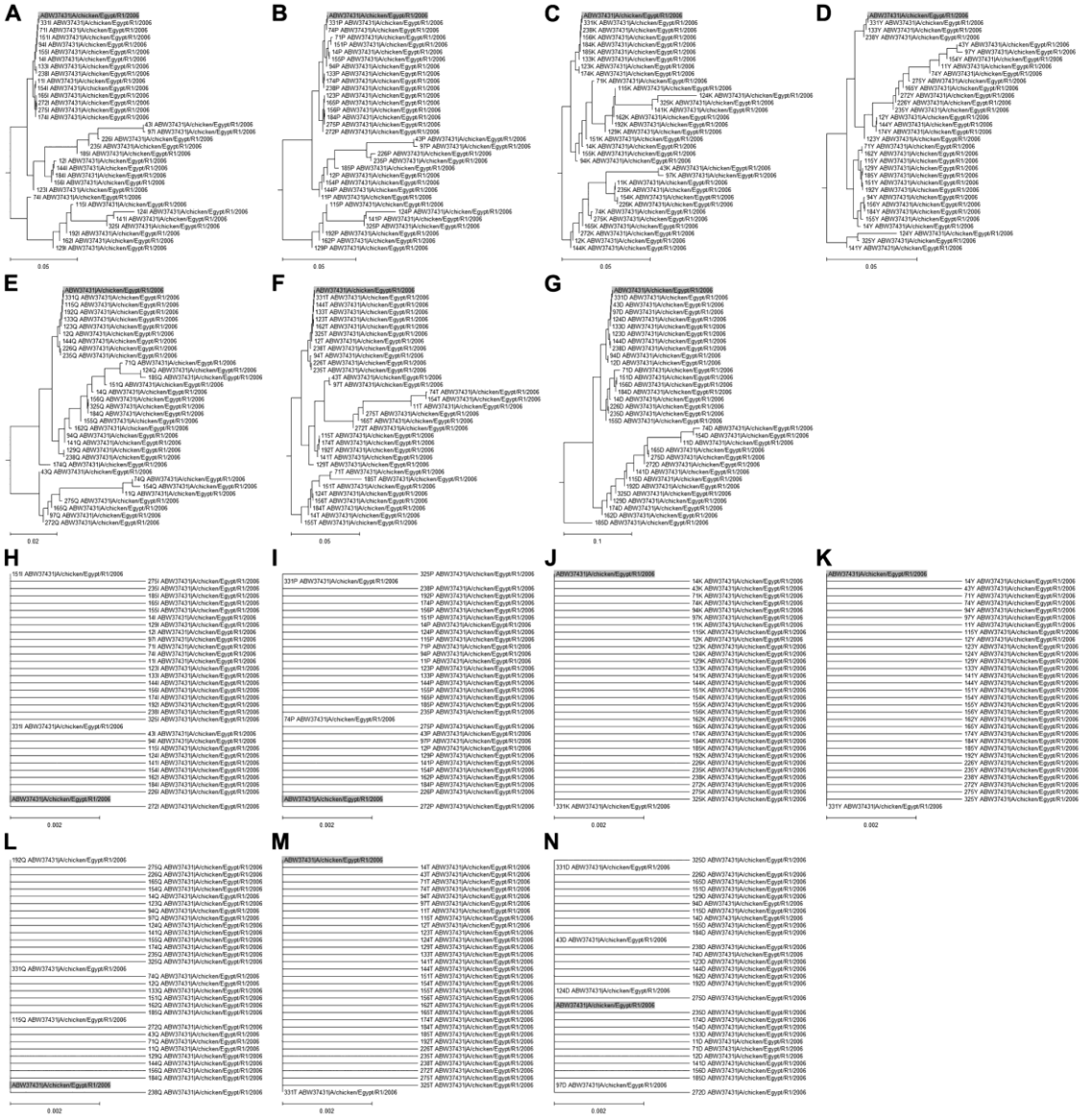
Rastojanje	**Amino-kiselina**	**Srednja vrednost**	**Standardno odstupanje**	**Koeficijent varijacije**
ISM	I	0.015968987	0.016462359	1.030895669
	A	0.014680109	0.007753097	0.528136207
	D	0.040545540	0.017047781	0.420460072
JTT	I	0.010035816	0.001381754	0.137682230
	A	0.009192425	0.002920772	0.317736800
	D	0.009981375	0.001367115	0.136966642

Tabela 4.1.2. Srednje vrednosti i standardna odstupanja ISM rastojanja odnosa amplituda na frekvencama F(0.236) i F(0.076), između HA1 proteina iz H5N1-HPAIV (GenBank: ABW37431) i odgovarajućih mutiranih sekvenci, uvođenjem pojedinačne supstitucije redom na svaku od 32 nekonzervirane pozicije proteina HA1, aminokiselinama I, P, K, Y, Q, T i D.

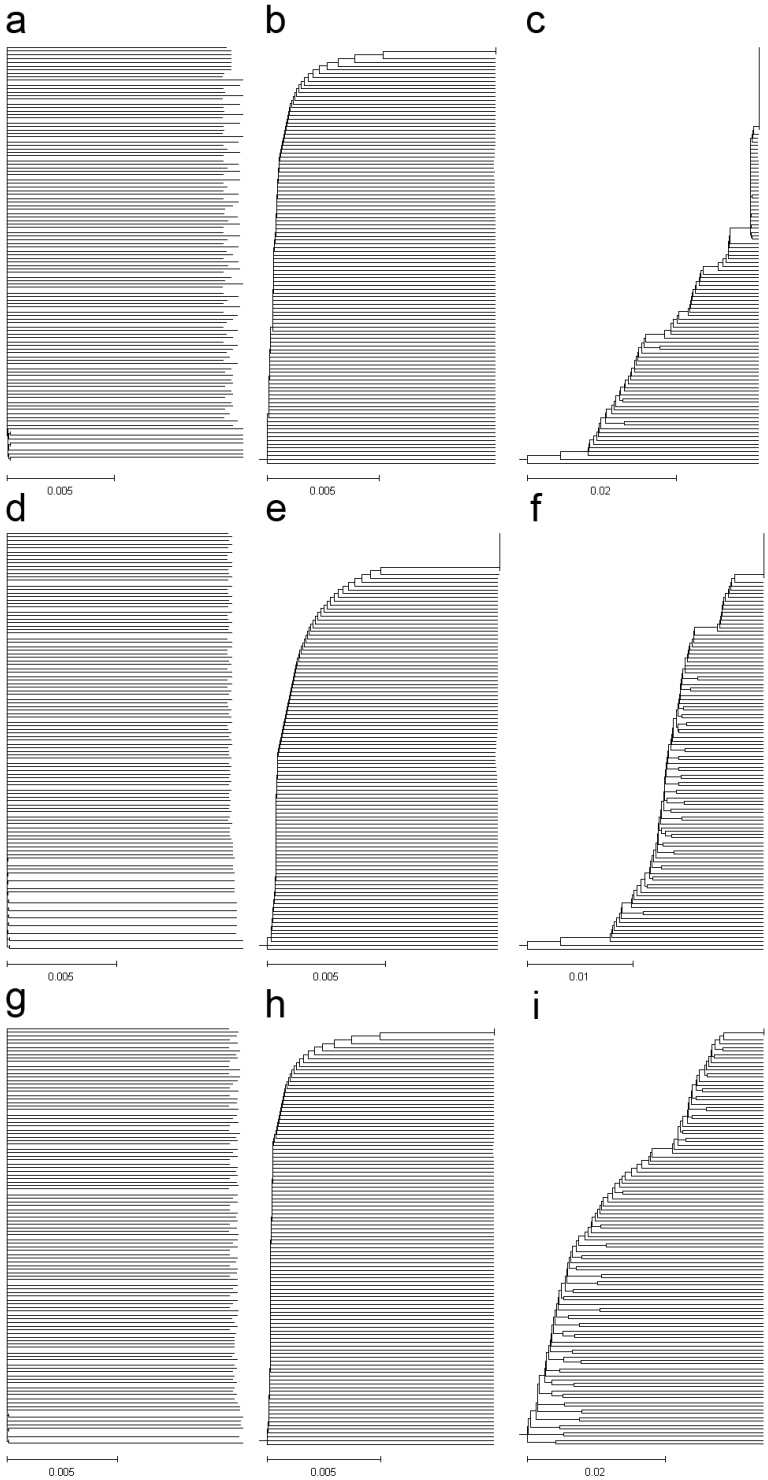
Amino-kiselina	**Srednja vrednost**	**Standardno odstupanje**	**Koeficijent varijacije**
I	0. 510654458	0. 050169727	0. 098245939
P	0. 51199869	0. 040066798	0. 078255665
K	0. 513499517	0. 033442625	0. 065126888
Y	0. 514971871	0. 030692419	0. 059600185
Q	0. 517952601	0. 033572762	0. 064818212
T	0. 520539452	0. 040824044	0. 078426417
D	0. 526009115	0. 059402374	0. 112930312



Slika 4.1.2. Osetljivost ISM rastojanja A(0.236)/A(0.076) na pojedinačne mutacije u nekonzerviranim pozicijama H5N1-HPAIV HA1 (GenBank: ABV37431). ISM rastojanje A(0.236)/A(0.076) je predstavljeno kao funkcija pojedinačnih supstitucija aminokiselinama I, P, K, Y, Q, T i D, na svakoj od 32 nekonzervirane pozicije proteina HA1.



Slika 4.1.3. Uticaj pozicije i tipa mutacije u HA1 iz H5N1-HPAIV, na filogenetsko izvođenje zasnovano na MSA i na ISM pristupu. Filogenetska stabla, za svaki skup od 32 sekvence dobijene uvođenjem pojedinačne aminokiseline (I, P, K, Y, Q, T i D) na svaku od 32 nekonzervirane pozicije HA1 proteina, su generisana primenom: (a)-(g) ISM rastojanja A(0.236)/A(0.076), (h)-(n) standardnim NJ metodom. Divlji tip (ABW37431/A/chicken/Egypt/R1/2006) je označen sivom bojom u svakom stablu. U sve ostale sekvence uvedena je pojedinačna supstitucija kiselinom: (a),(h) supstitucija aminokiselinom I, (b),(i) P, (c),(j) K, (d),(k) Y, (e),(l) Q, (f),(m) T, (g),(n) D.



Slika 4.1.4. Filogenetsko stablo za svaki skup od 110 sekvenci generisanih uvođenjem svake pojedinačne aminokiseline I, A, D, na svaku od 110 pozicija u humanom insulinu. Filogenetska stabla konstruisana primenom (a),(d),(g) standardnog NJ metoda zasnovanog na MSA, (b),(e),(h) UPGMA metoda zasnovanog na MSA, (c),(f),(i) algoritma zasnovanog na ISM rastojanju na celom spektru.

4.1.3. ISTREE algoritam

ISM filogenetsko stablo se generiše sledećim algoritmom:

1. Za svaku sekvencu izračunati njen informacioni spektar:
   1.1. Prevesti sekvencu aminokiselina u signal EIIP vrednosti.
   1.2. Smanjiti signal na srednju vrednost nule.
   1.3. Produžiti signal nulama do dužine najduže sekvence, kako bi se postavile rezolucije svih spektara na jednake vrednosti.
   1.4. Primenom brze Furijeove transformacije (FFT) na EIIP signalu, generisati informacioni spektar gustine energije.
2. Izračunati matricu rastojanja, koristeći metriku definisanu u (4.1), (4.2) ili (4.3)
3. Konstruisati stablo koristeći UPGMA ili NJ algoritam.

4.1.4. Kompleksnost ISTREE algoritma

Ukupna vremenska kompleksnost algoritma za generisanje ISM stabla je u najboljem slučaju $O(N(N+L\log L))$, a u najgorem slučaju $O(N(L\log L+N(N+L)))$, u zavisnosti od izbora vrste metrike i metode klasterisanja, gde je N broj sekvenci i L dužina najduže sekvence.

Za prvi korak algoritma vremenska složenost je $O(NL\log L))$, prema vremenskoj složenosti algoritma FFT koja iznosi $O(L\log L)$ [167]. Za izračunavanje matrice rastojanja u drugom koraku kada se koriste metrike d_1 i d_2 složenost je $O(N^2)$, a u slučaju metrike d_3 je $O(N^2L)$. U trećem koraku složenost klasterisanja metodom UPGMA je $O(N^2)$ [196], odnosno $O(N^3)$ za NJ algoritam [28].

4.1.5. Vreme računanja za ISTREE

Test performansi ISM filogenetskog algoritma je izvršen na simuliranom skupu sekvenci. U tabeli 4.1.3 su dati količina iskorišćene memorije i vremena računanja programa *ISMStablo*, u zavisnosti od broja i dužina proteinskih sekvenci. Korišćeno je rastojanje na celom spektru d_3 i UPGMA metoda. Skupovi proteinskih sekvenci su generisani principom slučajnog izbora, gde je broj sekvenci u intervalu od 100 do 4000,

a dužina sekvenci u opsegu od 100 do 10000. Testiranje je izvršeno i izmereno na računaru PC Pentium Dual-Core CPU E5200 2.50 GHz 3 GB RAM, na Windows XP operativnom sistemu. Za skup od 4000 sekvenci, gde su sve sekvence dužine 10000 aminokiselina, ukupno vreme računanja je 34 minuta i 17 sekundi. Detaljna vremena računanja za svaku fazu metode su data u tabeli 4.1.3 u formatu SP+DM+CL, gde je SP izmereno vreme za prvu fazu računanja informacionih spektara, DM za drugu fazu generisanja matrice rastojanja i CL za treću fazu klasterisanja.

Tabela 4.1.3. Vremena računanja i količina iskorišćene memorije programa ISMStablo na slučajno generisanim proteinskim sekvencama. Računar: PC Pentium Dual-Core CPU E5200 2.50 GHz 3 GB RAM. Sistem: Windows XP. Detaljna vremena za svaku fazu su data u formatu SP+DM+CL, gde je SP izmereno vreme za prvu fazu računanja informacionih spektara, DM za drugu fazu generisanja matrice rastojanja i CL za treću fazu klasterisanja.

Broj sekvenci	**Dužina sekvenci**	**Ukupno vreme**	**SP+DM+CL (sekunde)**	**Iskorišćena memorija (MB)**
100	100	0.11 sec	0.001+0.094+0.015	1
	1000	0.22 sec	0.047+0.158+0.015	2
	10000	2.844 sec	1.610+1.219+0.015	9
1000	100	11.06 sec	0.06+9+2	18
	1000	17.5 sec	0.5+15+2	20
	10000	2 min 16 sec	16+118+2	74
2000	100	50.12 sec	0.12+35+15	35
	1000	1 min 15 sec	1.1+59+15	74
	10000	9 min	27+498+15	193
4000	100	4 min 22 sec	0.26+141+121	254
	1000	5 min 58 sec	2.17+235+121	269
	10000	34 min 17 sec	53+1880+121	499

4.1.6. Testiranje ISM filogenetskog pristupa

ISM filogenetski pristup je testiran na primerima različitih važnih klasa ćelijskih proteina: (i) glukokortikoidnom receptoru (nuklearnom receptoru) koji ima važnu ulogu u razvoju organizma, metabolizmu i imunskom odgovoru, (ii) hormonu leptinu koji ima ulogu u regulaciji unosa i potrošnje energije, (iii) insulinu, glavnom hormonu za regulaciju metabolizma ugljenih hidrata i masti, i (iv) enzimu lipoproteinskoj lipazi (LPL) koja ima ključnu ulogu u razlaganju triglicerida. Sve kompletne sekvence glukokordikoidnog receptora, insulina, leptin hormona i enzima lipoprotein lipaze (LPL) su preuzete iz NCBI baze [138] i upotrebljene za upoređivanje ISM filogenetskog pristupa sa standardnim metodama.

Od programskih alata, za standardne filogenetske metode zasnovane na rastojanju, korišćen je programski paket MEGA5 [197], a za izvođenje ML stabla je primenjen PHYML alat [46]. Za izračunavanje MSA sekvenci korišćen je MUSCLE algoritam [198] u programu MEGA5.

Za testiranje i upoređivanje standardnih i ISM filogenetskih pristupa, za svaki od sledećih grupa sekvenci:

(i) glukokortikoidni receptor
(ii) protein insulin
(iii) hormon leptin
(iv) enzim lipoproteinska lipaza

generisana su filogenetska stabla korišćenjem:

(a) ML metode sa NNI pretraživanjem topologije stabala i BioNJ metodom početnog pretraživanja
(b) UPGMA metode sa „Poisson“ modelom korekcije [197], gde su sve dvosmislene pozicije uklonjene iz svakog para sekvenci
(c) ISM pristupa sa merom rastojanja na celom spektru

Upoređivanje ISM pristupa sa standardnim filogenetskim metodama pokazuje sličnu biološku klasifikaciju taksonomskih klasa i familija vrsta organizama, ali takođe otkriva određene razlike (slika 4.1.5, slika 4.1.6). Ove razlike nisu značajne i odnose se na pomeranja unutar biološki povezanih grana. Na primer:

(i) Samo u ISM stablima glukortikoidnog receptora i proteina leptina, sve vrste glodara su grupisane zajedno (slika 4.1.5c,i)

(ii) U ISM i UPGMA stablima enzima LPL, gvinejsko prase je izdvojeno od glodara (slika 4.1.5k,l)

(iii) U ML stablu leptina zec je pogrešno pozicioniran u grani primata (slika 4.1.5i) itd.

Sva stabla na slici 4.1.5 prate standardnu klasifikaciju na: *Eutheria* (placentalni sisari), *Marsupialia* (torbari), *Aves* (ptice), *Salientia* (žabe) i *Teleostei* (ribe).

Algoritmi za standardnu analizu u molekularnoj filogenezi, koji koriste informacije dobijene strukturnom analizom određenih grupa gena u različitim vrstama organizama, su principijelno različiti od filogenetskog algoritma zasnovanog na ISM metodi koji analizira evoluciju biološke funkcije jednog gena kroz više vrsta. Upoređivanje stabala na slici 4.1.5, sa opšte prihvaćenim molekularno filogenetskim stablima placentalnih sisara [199-201], prikazuje slično klasterisanje placentalnih sisara na *Afrotheria*, *Eutheria*, *Laurasiatheria*, sa nekim razlikama:

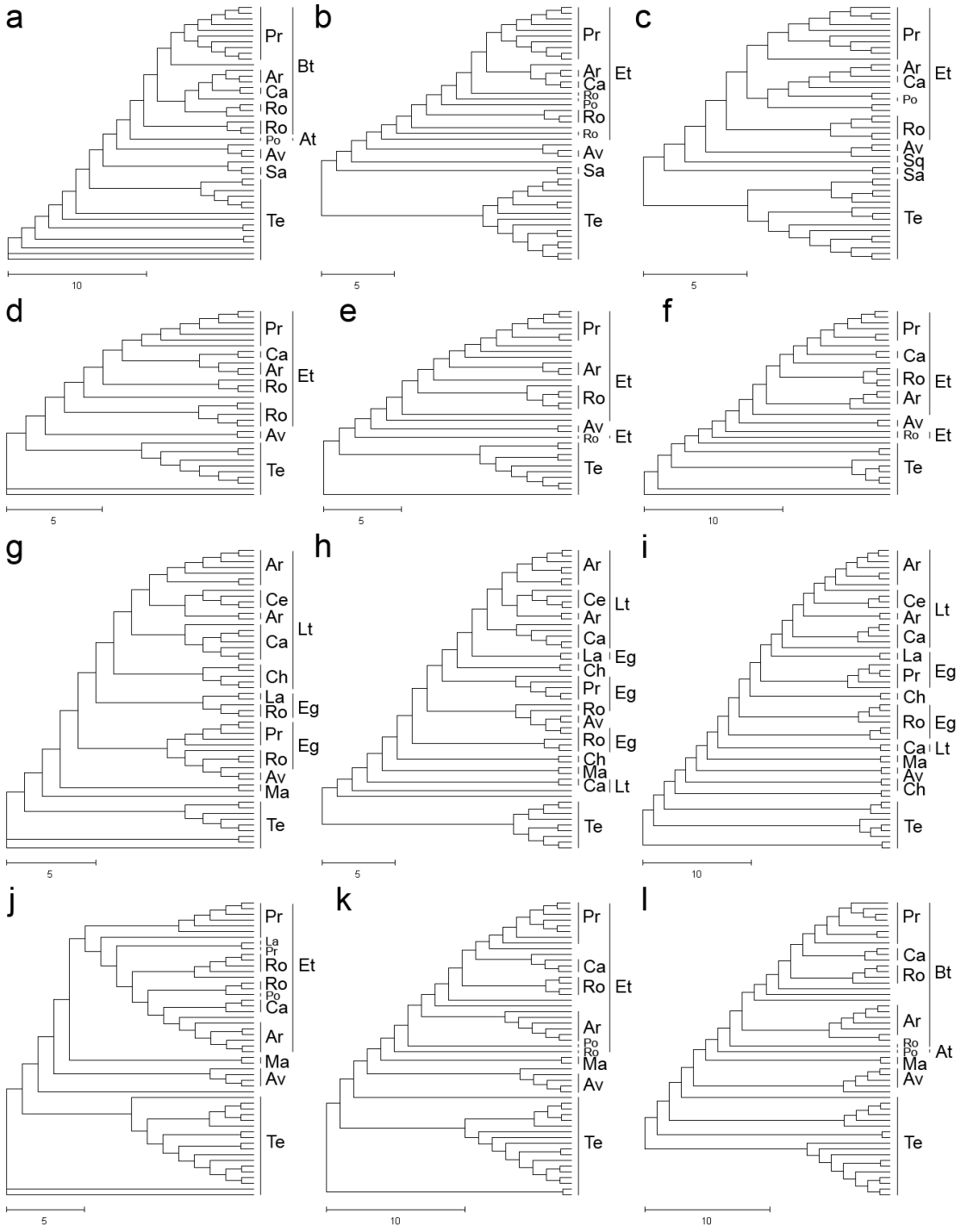
(i) Za glukokortikoidni receptor (slika 4.1.5a-c) *Loxodonta Africana* (*Afrotheria*) je odvojen od *Boreoeutheria* grupe samo u ML stablu (slika 4.1.5a).

(ii) U slučaju proteina insulina (slika 4.1.5d-f), sve vrste *Eutheria* su grupisane zajedno u ML stablu, dok su u UPGMA I ISM stablu nekoliko glodara odvojeni od grane *Eutheria*.

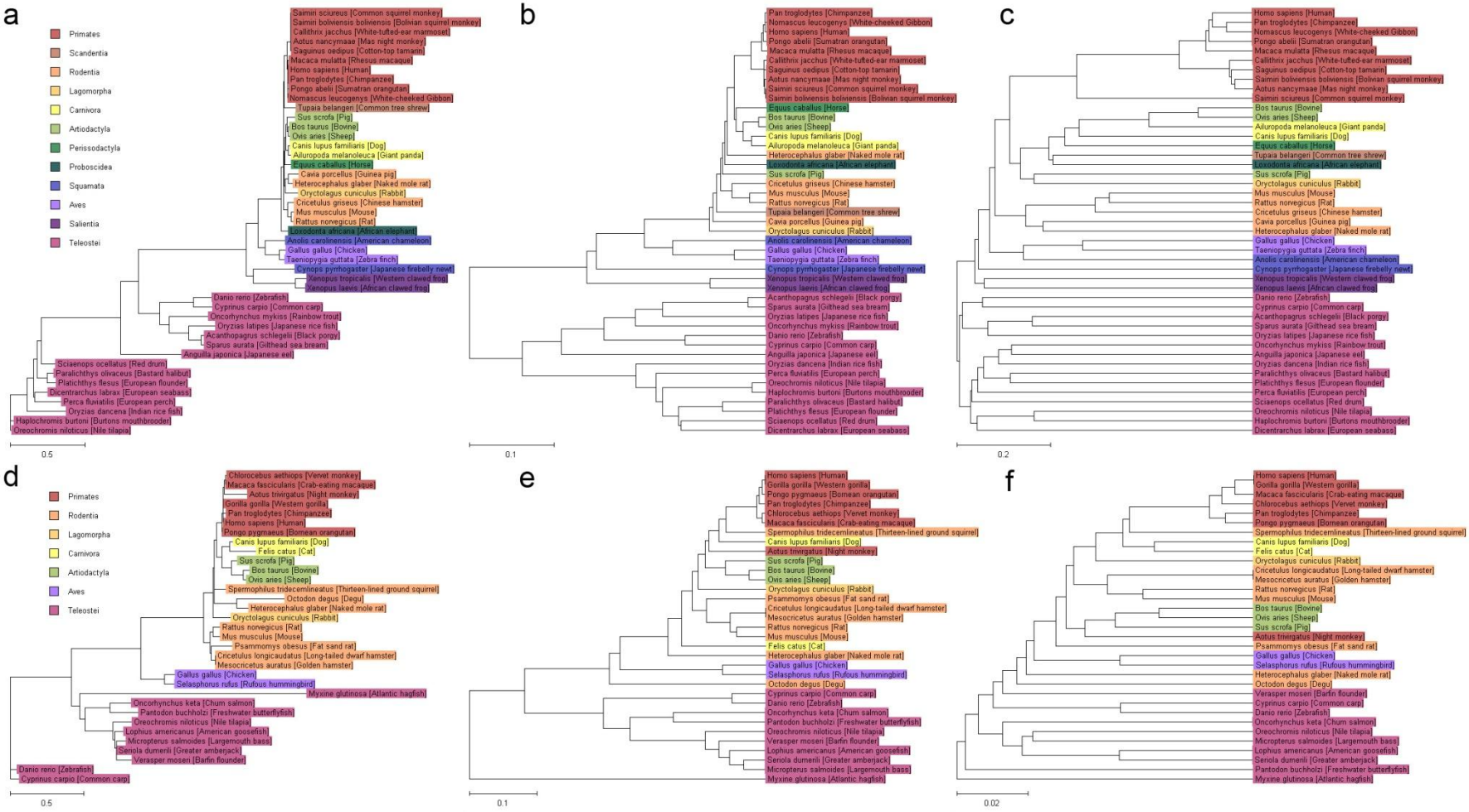
(iii) Stabla hormona leptina (slika 4.1.5g-i) pokazuje najviše sličnosti sa standardnim molekularno filogenetskim stablom, jedino su nekoliko *Chiroptera* (slepi miševi) u ISM i UPGMA stablima odvojeni od *Eutheria* klastera (slika 4.1.5h,i), i u UPGMA stablu (slika 4.1.5h) *Aves* (ribe) su pogrešno grupisane u *Eutheria* grani.

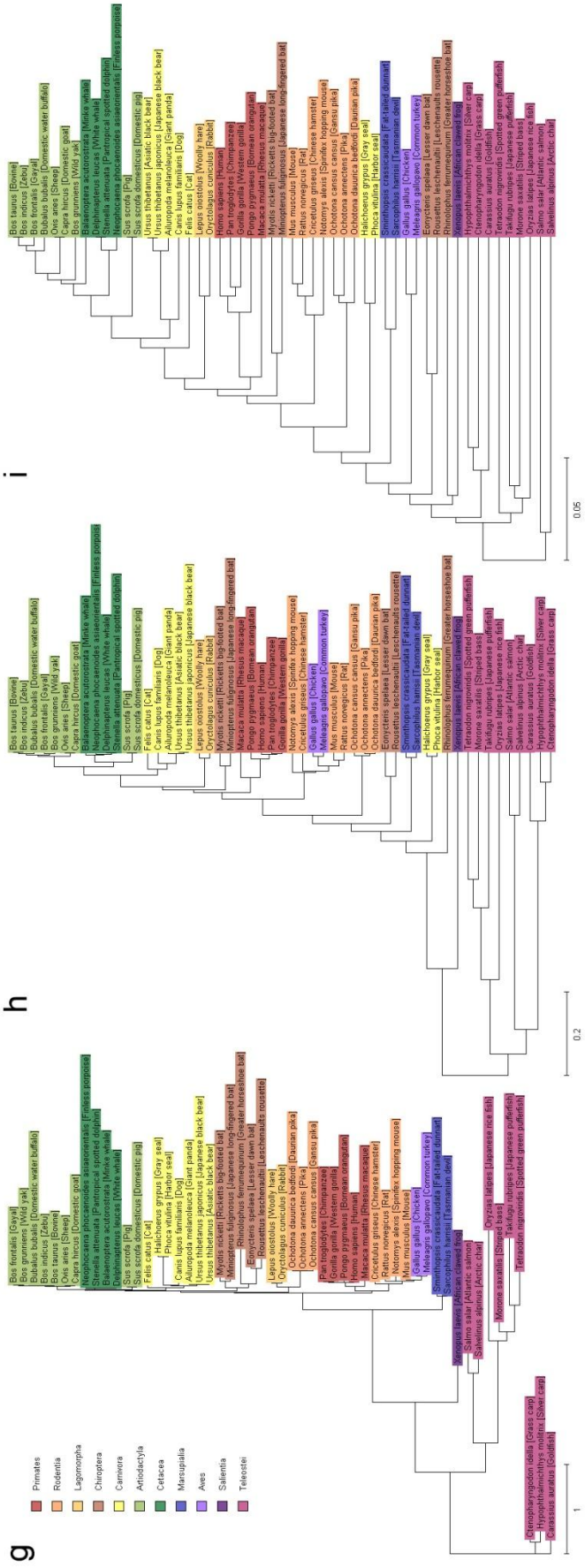
(iv) U LPL stablima (slika 4.1.5j-l) *Afrotheria* je odvojena od *Boreoeutheria* skupa jedino u ISM stablu (slika 4.1.5l).

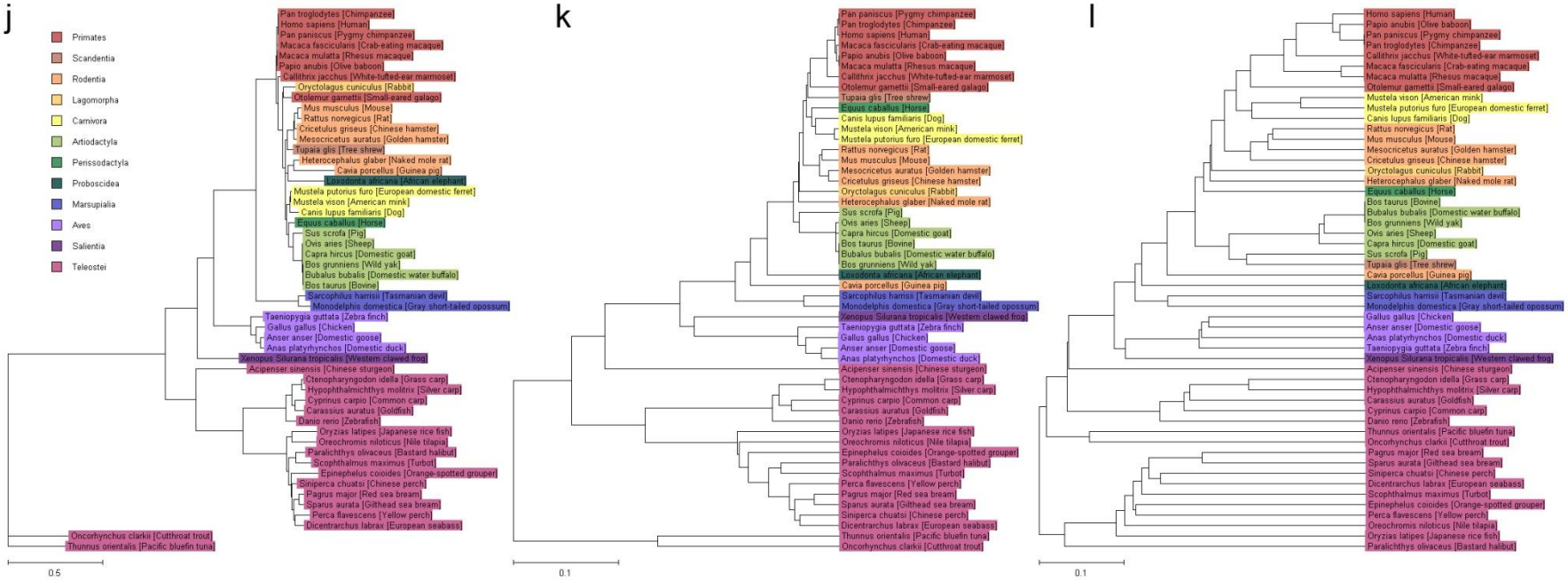
Na slici 4.1.5 su korišćene sledeće skraćenice za taksonomske klase i familije vrsta organizama: *Ar* Artiodactyla, *Av* Aves, *Ca* Carnivora, *Ce* Cetacea, *Ch* Chiroptera, *La* Lagomorpha, *Ma* Marsupialia, *Pr* Primates, *Po* Proboscidea, *Ro* Rodentia, *Sa* Salientia, *Sc* Scandentia, *Sq* Squamata, *Te* Teleostei. Skraćenice za klade: *At* Afrotheria, *Bt* Boreoeutheria, *Eg* Euarchontoglires, *Et* Eutheria, *Lt* Laurasiatheria.



Slika 4.1.5. Upoređivanje standardnih metoda sa ISM filogenetskom analizom. Filogenetsko stablo konstruisano primenom: (a),(d),(g),(j) standardne ML metode; (b),(e),(h),(k) standardne UPGMA metode; (c),(f),(i),(l) ISM metode; za: (a),(b),(c) glukokortikoidni receptor; (d)-(f) protein insulin; (g)-(i) hormon leptin; (j)-(l) enzim lipoprotein lipaze.


a
Primates
Scandentia
Rodentia
Lagomorpha
Carnivora
Artiodactyla
Perissodactyla
Proboscidea
Squamata
Aves
Salientia
Teleostei
Saimiri sciureus [Common squirrel monkey]
Saimiri boliviensis boliviensis [Bolivian squirrel monkey]
Callithrix jacchus [White-tufted-ear marmoset]
Aotus nancymaae [Mas night monkey]
Saguinus oedipus [Cotton-top tamarin]
Macaca mulatta [Rhesus macaque]
Homo sapiens [Human]
Pan troglodytes [Chimpanzee]
Pongo abelii [Sumatran orangutan]
Nomascus leucogenys [White-cheeked Gibbon]
Tupaia belangeri [Common tree shrew]
Sus scrofa [Pig]
Bos taurus [Bovine]
Ovis aries [Sheep]
Canis lupus familiaris [Dog]
Ailuropoda melanoleuca [Giant panda]
Equus caballus [Horse]
Cavia porcellus [Guinea pig]
Heterocephalus glaber [Naked mole rat]
Oryctolagus cuniculus [Rabbit]
Cricetulus griseus [Chinese hamster]
Mus musculus [Mouse]
Rattus norvegicus [Rat]
Loxodonta africana [African elephant]
Anolis carolinensis [American chameleon]
Gallus gallus [Chicken]
Taeniopygia guttata [Zebra finch]
Cynops pyrrhogaster [Japanese firebelly newt]
Xenopus tropicalis [Western clawed frog]
Xenopus laevis [African clawed frog]
Danio rerio [Zebrafish]
Cyprinus carpio [Common carp]
Oncorhynchus mykiss [Rainbow trout]
Oryzias latipes [Japanese rice fish]
Acanthopagrus schlegelii [Black porgy]
Sparus aurata [Gilthead sea bream]
Anguilla japonica [Japanese eel]
Sciaenops ocellatus [Red drum]
Paralichthys olivaceus [Bastard halibut]
Platichthys flesus [European flounder]
Dicentrarchus labrax [European seabass]
Perca fluviatilis [European perch]
Oryzias dancena [Indian rice fish]
Haplochromis burtoni [Burtons mouthbrooder]
Oreochromis niloticus [Nile tilapia]
0.5
b
Pan troglodytes [Chimpanzee]
Nomascus leucogenys [White-cheeked Gibbon]
Homo sapiens [Human]
Pongo abelii [Sumatran orangutan]
Macaca mulatta [Rhesus macaque]
Callithrix jacchus [White-tufted-ear marmoset]
Saguinus oedipus [Cotton-top tamarin]
Aotus nancymaae [Mas night monkey]
Saimiri sciureus [Common squirrel monkey]
Saimiri boliviensis boliviensis [Bolivian squirrel monkey]
Equus caballus [Horse]
Bos taurus [Bovine]
Ovis aries [Sheep]
Canis lupus familiaris [Dog]
Ailuropoda melanoleuca [Giant panda]
Heterocephalus glaber [Naked mole rat]
Loxodonta africana [African elephant]
Sus scrofa [Pig]
Cricetulus griseus [Chinese hamster]
Mus musculus [Mouse]
Rattus norvegicus [Rat]
Tupaia belangeri [Common tree shrew]
Cavia porcellus [Guinea pig]
Oryctolagus cuniculus [Rabbit]
Anolis carolinensis [American chameleon]
Gallus gallus [Chicken]
Taeniopygia guttata [Zebra finch]
Cynops pyrrhogaster [Japanese firebelly newt]
Xenopus tropicalis [Western clawed frog]
Xenopus laevis [African clawed frog]
Acanthopagrus schlegelii [Black porgy]
Sparus aurata [Gilthead sea bream]
Oryzias latipes [Japanese rice fish]
Oncorhynchus mykiss [Rainbow trout]
Danio rerio [Zebrafish]
Cyprinus carpio [Common carp]
Anguilla japonica [Japanese eel]
Oryzias dancena [Indian rice fish]
Perca fluviatilis [European perch]
Oreochromis niloticus [Nile tilapia]
Haplochromis burtoni [Burtons mouthbrooder]
Paralichthys olivaceus [Bastard halibut]
Platichthys flesus [European flounder]
Sciaenops ocellatus [Red drum]
Dicentrarchus labrax [European seabass]
0.1
c
Homo sapiens [Human]
Pan troglodytes [Chimpanzee]
Nomascus leucogenys [White-cheeked Gibbon]
Pongo abelii [Sumatran orangutan]
Macaca mulatta [Rhesus macaque]
Callithrix jacchus [White-tufted-ear marmoset]
Saguinus oedipus [Cotton-top tamarin]
Saimiri boliviensis [Bolivian squirrel monkey]
Aotus nancymaae [Mas night monkey]
Saimiri sciureus [Common squirrel monkey]
Bos taurus [Bovine]
Ovis aries [Sheep]
Ailuropoda melanoleuca [Giant panda]
Canis lupus familiaris [Dog]
Equus caballus [Horse]
Tupaia belangeri [Common tree shrew]
Loxodonta africana [African elephant]
Sus scrofa [Pig]
Oryctolagus cuniculus [Rabbit]
Mus musculus [Mouse]
Rattus norvegicus [Rat]
Cricetulus griseus [Chinese hamster]
Cavia porcellus [Guinea pig]
Heterocephalus glaber [Naked mole rat]
Gallus gallus [Chicken]
Taeniopygia guttata [Zebra finch]
Anolis carolinensis [American chameleon]
Cynops pyrrhogaster [Japanese firebelly newt]
Xenopus tropicalis [Western clawed frog]
Xenopus laevis [African clawed frog]
Danio rerio [Zebrafish]
Cyprinus carpio [Common carp]
Acanthopagrus schlegelii [Black porgy]
Sparus aurata [Gilthead sea bream]
Oryzias latipes [Japanese rice fish]
Oncorhynchus mykiss [Rainbow trout]
Anguilla japonica [Japanese eel]
Oryzias dancena [Indian rice fish]
Paralichthys olivaceus [Bastard halibut]
Platichthys flesus [European flounder]
Perca fluviatilis [European perch]
Sciaenops ocellatus [Red drum]
Oreochromis niloticus [Nile tilapia]
Haplochromis burtoni [Burtons mouthbrooder]
Dicentrarchus labrax [European seabass]
0.2
d
Primates
Rodentia
Lagomorpha
Carnivora
Artiodactyla
Aves
Teleostei
Chlorocebus aethiops [Vervet monkey]
Macaca fascicularis [Crab-eating macaque]
Aotus trivirgatus [Night monkey]
Gorilla gorilla [Western gorilla]
Pan troglodytes [Chimpanzee]
Homo sapiens [Human]
Pongo pygmaeus [Bornean orangutan]
Canis lupus familiaris [Dog]
Felis catus [Cat]
Sus scrofa [Pig]
Bos taurus [Bovine]
Ovis aries [Sheep]
Spermophilus tridecemlineatus [Thirteen-lined ground squirrel]
Octodon degus [Degu]
Heterocephalus glaber [Naked mole rat]
Oryctolagus cuniculus [Rabbit]
Rattus norvegicus [Rat]
Mus musculus [Mouse]
Psammomys obesus [Fat sand rat]
Cricetulus longicaudatus [Long-tailed dwarf hamster]
Mesocricetus auratus [Golden hamster]
Gallus gallus [Chicken]
Selasphorus rufus [Rufous hummingbird]
Myxine glutinosa [Atlantic hagfish]
Oncorhynchus keta [Chum salmon]
Pantodon buchholzi [Freshwater butterflyfish]
Oreochromis niloticus [Nile tilapia]
Lophius americanus [American goosefish]
Micropterus salmoides [Largemouth bass]
Seriola dumerili [Greater amberjack]
Verasper moseri [Barfin flounder]
Danio rerio [Zebrafish]
Cyprinus carpio [Common carp]
0.5
e
Homo sapiens [Human]
Gorilla gorilla [Western gorilla]
Pongo pygmaeus [Bornean orangutan]
Pan troglodytes [Chimpanzee]
Chlorocebus aethiops [Vervet monkey]
Macaca fascicularis [Crab-eating macaque]
Spermophilus tridecemlineatus [Thirteen-lined ground squirrel]
Canis lupus familiaris [Dog]
Aotus trivirgatus [Night monkey]
Sus scrofa [Pig]
Bos taurus [Bovine]
Ovis aries [Sheep]
Oryctolagus cuniculus [Rabbit]
Psammomys obesus [Fat sand rat]
Cricetulus longicaudatus [Long-tailed dwarf hamster]
Mesocricetus auratus [Golden hamster]
Rattus norvegicus [Rat]
Mus musculus [Mouse]
Felis catus [Cat]
Heterocephalus glaber [Naked mole rat]
Gallus gallus [Chicken]
Selasphorus rufus [Rufous hummingbird]
Octodon degus [Degu]
Cyprinus carpio [Common carp]
Danio rerio [Zebrafish]
Oncorhynchus keta [Chum salmon]
Pantodon buchholzi [Freshwater butterflyfish]
Oreochromis niloticus [Nile tilapia]
Verasper moseri [Barfin flounder]
Lophius americanus [American goosefish]
Seriola dumerili [Greater amberjack]
Micropterus salmoides [Largemouth bass]
Myxine glutinosa [Atlantic hagfish]
0.1
f
Homo sapiens [Human]
Gorilla gorilla [Western gorilla]
Macaca fascicularis [Crab-eating macaque]
Chlorocebus aethiops [Vervet monkey]
Pan troglodytes [Chimpanzee]
Pongo pygmaeus [Bornean orangutan]
Spermophilus tridecemlineatus [Thirteen-lined ground squirrel]
Canis lupus familiaris [Dog]
Felis catus [Cat]
Oryctolagus cuniculus [Rabbit]
Cricetulus longicaudatus [Long-tailed dwarf hamster]
Mesocricetus auratus [Golden hamster]
Rattus norvegicus [Rat]
Mus musculus [Mouse]
Bos taurus [Bovine]
Ovis aries [Sheep]
Sus scrofa [Pig]
Aotus trivirgatus [Night monkey]
Psammomys obesus [Fat sand rat]
Gallus gallus [Chicken]
Selasphorus rufus [Rufous hummingbird]
Heterocephalus glaber [Naked mole rat]
Octodon degus [Degu]
Verasper moseri [Barfin flounder]
Cyprinus carpio [Common carp]
Danio rerio [Zebrafish]
Oncorhynchus keta [Chum salmon]
Oreochromis niloticus [Nile tilapia]
Micropterus salmoides [Largemouth bass]
Lophius americanus [American goosefish]
Seriola dumerili [Greater amberjack]
Pantodon buchholzi [Freshwater butterflyfish]
Myxine glutinosa [Atlantic hagfish]
0.02


g
h
i
Primates
Rodentia
Lagomorpha
Chiroptera
Carnivora
Artiodactyla
Cetacea
Marsupialia
Aves
Salentia
Teleostei
1
0.2
0.05



Slika 4.1.6. Upoređivanje standardnih metoda sa ISM zasnovanom filogenetskom analizom. Stabla generisana kao na slici 4.1.5. prikazana su u visokoj rezoluciji i sa više detalja.

4.2. Bioinformatička platforma zasnovana na EIIP/ISM

Za analizu protein-protein interakcija, analizu veze između strukture i biološke funkcije proteina, procenu biološkog efekta mutacija, dizajniranje proteina i peptida željene biološke aktivnosti i funkcionalnu filogenetsku analizu, sa tim da su analize bazirane na modelu dalekodosežnih međumolekulskih interakcija, bilo je neophodno razviti softverski paket. Kako potencijal elektron-jon interakcije (EIIP) predstavlja osnovni fizički parametar molekula u dalekodosežnim interakcijama, a čini osnovu ISM metode, razvijen je softverski paket koji je zasnovan na parametru EIIP i metodi ISM i nazvan je *Bioinformatička platforma zasnovana na EIIP/ISM*, ili kraće *EIIP/ISM platforma*. Platforma je podeljena u osam modula i sastoji se od 31 programskog alata.

Zbog komplikovanosti algoritama ISM metode, velike količine podataka u bazama koja se analiziraju i dužina sekvenci od preko hiljadu aminokiselina, kritičan resurs je brzina kojom se izvršavaju algoritmi. Platforma je razvijena u *C++* programskom jeziku, u *Microsoft Visual Studio 6.0* razvojnom okruženju. Ukupna veličina izvornog koda iznosi oko 1.9 MB.

4.2.1. Osnova platforme

Jezgro platforme čine metode za izračunavanje informacionog spektra i konsenzus informacionog spektra, odnosno kros-spektra.

4.2.1.1. Algoritam za izračunavanje informacionog spektra

Osnovni algoritam ISM metode je generisanje informacionog spektra proteinske sekvence, koji se sastoji iz dva osnovna koraka: (i) transformacija primarne strukture u signal kodiranjem svake aminokiseline EIIP vrednošću i (ii) Furijeova transformacija EIIP signala u spektar.

Ulaz: Proteinska sekvenca

Izlaz: Informacioni spektar

1) EIIP kodiranje sekvence u signal

Primarna struktura proteina dužine *N* se kodira u diskretni signal, tako sto se svaki element (aminokiselinski ostatak) kodira u EIIP vrednost. EIIP je osobina aminokiselina i nukleotida, koja je zadužena za dalekodosežne međumolekulske interakcije.

2) Spuštanje signala na nultu srednju vrednost

   Ceo signal se umanjuje za vrednost srednje vrednosti signala, tako da novodobijeni signal ima srednju vrednost nula. Ovim procesom se eliminiše visoka amplituda na nultoj frekvenci (što može izazvati prekoračenje pri izvršenju Furijeove transformacije signala velike dužine) koja je posledica srednje vrednosti signala iznad nule i može se tumačiti kao komponenta signala s beskonačno velikim periodom oscilovanja. Ostatak spektra ostaje nepromenjen u odnosu na spektar originalnog signala.

3) Produžavanje signala nulama do potrebne dužine

   Signal se produžava nulama do najmanjeg stepena dvojke u slučaju FFT transformacije, a u slučaju kros-spektra i do dužine najdužeg signala. Procesom produžavanja nulama (eng. *zero padding*) se dobija spektar veće rezolucije.

4) Furijeova diskretna transformacija signala u spektar

   Primenom brza Furijeove transformacije (eng. *Fast Fourier Transform, FFT*) ili diskretne Furijeove transformacije (eng. *Discrete Fourier Transform, DFT*) izračunava se spektar signala

5) Generisanje informacionog spektra

   Spektar gustine energije se dobija množenjem spektra sa svojim konjugatom na prostoru frekvenci.

4.2.1.2. Algoritam za izračunavanje informacionog kros-spektra

Za identifikaciju karakterističnih frekventnih komponenti informacionih spektara grupe proteina, koje su značajne za određenu biološku funkciju, vrši se konsenzus spektralna analiza ISM metode primenom algoritma za izračunavanje informacionog kros-spektra.

Ulaz: niz proteinskih sekvenci

Izlaz: Informacioni krosspektar

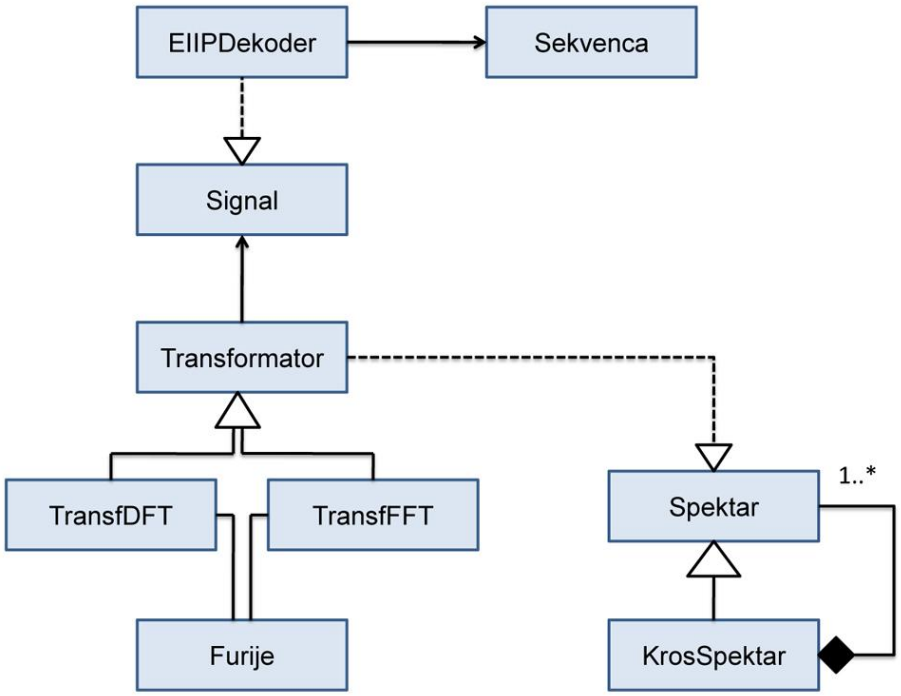
1) Za svaku sekvencu se izračuna informacioni spektar kao u osnovnom algoritmu, uz prethodno produžavanje svih signala do najmanjeg stepena dvojke najduže sekvence, kako bi svi spektri bili iste rezolucije i dužine za FFT algoritam (2^N)
2) Krosspektar se dobija pokoordinatnim množenjem svih spektara

4.2.1.3. Jezgro platforme

EIIP/ISM platforma je implementirana u objektno orijentisanom jeziku C++. Jezgro platforme čine objekti pomoću kojih se apstraktno opisuje ISM metoda. Osnovni objekti predstavljaju sekvencu, signal i spektar, a operacije koje se izvode nad njima su prevođenje sekvence u EIIP signal i transformacija signala u spektar. Na slici 4.2.1 je prikazan klasni dijagram odnosa osnovnih klasa u jezgru EIIP/ISM platforme.

Osnovne klase implementirane u jezgru platforme su:

- *Sekvenca*, koja sadrži zapis sekvence u obliku niza aminokiselinskih ostataka, sa nazivom i tipom sekvence (protein, RNK, DNK);
- *EIIPDekoder*, koji kreira objekat *Signal* od objekta *Sekvenca*, koristeći EIIP kodiranje svakog aminokiselinskog ostatka;
- *Signal*, gde je EIIP signal određene sekvence predstavljen nizom EIIP vrednosti.
- *Transformator*, koji ima funkciju kreiranja objekta *Spektar* za određeni objekat *Signal*, primenom FFT ili DFT transformacije;
- *Furije*, koji sadrži implementirane algoritme za Furijeove FFT i DFT transformacije niza kompleksnih brojeva;
- *Spektar*, u kojem je predstavljen informacioni spektar određene proteinske sekvence, zapisan nizovima amplituda, faza i nizom sortiranih pikova;
- *Krosspektar*, koji sadrži reference na spektre koji ga čine, a sam je podklasa klase *Spektar*. Klasa Krosspektar je implementirana kompozitnim šablonom (eng. *Composit pattern*).



Slika 4.2.1. Klasni dijagram jezgra ISM platforme.

Osnova metode generisanja informacionog spektra se zasniva na Furijeovoj transformaciji i njenim impementiranim algoritmima: diskretna Furijeova transformacija (DFT) i brza Furijeova transformacija (FFT).

Dat je primer izvornog koda za klasu *Furije* u C++ jeziku, sa implementiranim metodama za FFT i DFT transformaciju:

```
#include <complex>
using std::complex;
// ********* klasa FURIE *********
class Furie{
private:
static complex<double> Furie::root1(int a);
static void fft_gl(int duzinalog2, complex<double> *vPotencijal,
                        complex<double> *vSpektar, short idir);
static void dft_gl(int duzinaPotencijal, complex<double> *vPotencijal,
                        complex<double> *vSpektar, short idir);
public:
// Direktan FFT. Signal je duzine 2 na dlog2
static void fft(int dlog2, complex<double> *vUlaz,
      complex<double> *vIzlaz){ fft_gl(dlog2, vUlaz, vIzlaz, 0); }
// Inverzan FFT. Signal je duzine 2 na dlog2
static void fft_inv(int dlog2, complex<double> *vUlaz,
      complex<double> *vIzlaz){ fft_gl(dlog2, vUlaz, vIzlaz, 1); }
// Direktan DFT
static void dft(int duzina, complex<double> *vUlaz,
      complex<double> *vIzlaz){ dft_gl(duzina, vUlaz, vIzlaz, 0); }
```

```
// Inverzan DFT
static void dft_inv(int duzina, complex<double> *vUlaz,
      complex<double> *vIzlaz){ dft_gl(duzina, vUlaz, vIzlaz, 1); }
};

// Racuna spektar primenom FFT. Vektori su duzine 2 na duzinalog2.
// za idir==0 je direktna, inace inverzna.
void Furie::fft_gl(int duzinalog2, complex<double> *vPotencijal,
                        complex<double> *vSpektar, short idir){
      complex<double> t,u,w;
      int nv2,nm1,le,le1,i,j,l,k,ip,duzina=1;
      // duzina = 2 na duzinalog2
      for(i=0; i<duzinalog2; i++) duzina *= 2;
      nv2 = duzina/2;
      nm1 = duzina-1;
      j = 1;
      for(i=0; i<duzina; i++) vSpektar[i] = vPotencijal[i];
      for(i=1; i<=nm1; i++){
            if(i < j){
                  t = vSpektar[j-1];
                  vSpektar[j-1] = vSpektar[i-1];
                  vSpektar[i-1] = t;
            }
            for(k=nv2; k<j; k=k/2) j = j - k;
            j = j + k;
      }
      le = 1;
      l = 1;
      do{
            le = le * 2; // le=2 na l
            le1 = le / 2;
            u = complex<double>(1.0,0.0);
            if(idir==0) w = complex<double>(cos(pi/le1), sin(pi/le1));
            else w = complex<double>(cos(pi/le1), -sin(pi/le1));
            for(j=1; j<=le1; j++){
                  i = j;
                  do{
                        ip = i + le1;
                        t = vSpektar[ip-1] * u ;
                        vSpektar[ip-1] = vSpektar[i-1] - t;
                        vSpektar[i-1] = vSpektar[i-1] + t;
                        i = i + le;
                  }while(i <= duzina);
                  u = u * w;
            }
            l++;
      }while(l <= duzinalog2);
      if(idir == 1){
            for(j=0; j<duzina; j++) vSpektar[j] /= duzina;
      }
}
// Racuna a-ti kompleksan koren jedinice
complex<double> Furie::root1(int a){
      if(a == 0) throw "deljenje sa nulom";
      return complex<double>(cos(asin(1)*4/a),sin(asin(1)*4/a));
}
// Racuna spektar primenom DFT. idir==0 direktna, inace inverzna.
void Furie::dft_gl(int duzinaPotencijal, complex<double> *vPotencijal,
                        complex<double> *vSpektar, short idir){
```

```
	complex<double> w,ww,wm,s;
	if(idir == 0) w = root1(duzinaPotencijal);
	else w = root1(-duzinaPotencijal);
	int i,j;
	for(j=0,wm=1; j<duzinaPotencijal; j++,wm*=w){
		s = 0;
		for(i=0,ww=1; i<duzinaPotencijal; i++,ww*=wm)
			s+=(vPotencijal[i]*ww);
		vSpektar[j] = s;
	}
	if(idir==1)
		for(j=0; j<duzinaPotencijal; j++)
			vSpektar[j]/=duzinaPotencijal;
}
```

4.2.2. Struktura platforme

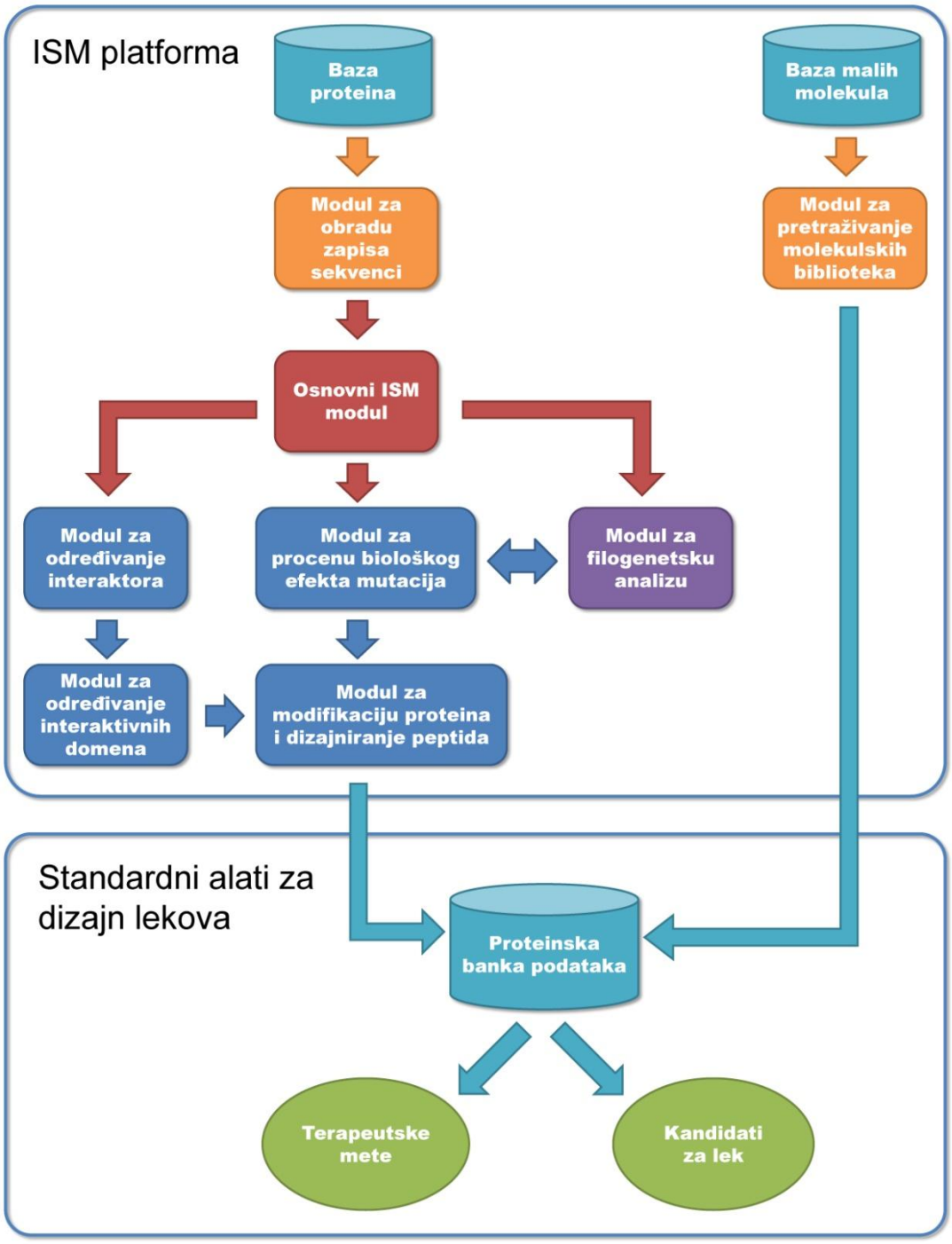
EIIP/ISM platforma se sastoji od osam modula. Svaki modul sadrži nekoliko programskih alata, pri čemu svaki ima određenu svrhu karakterističnu za taj modul (tabela 4.2.1).

Tabela 4.2.1. Spisak modula sa pripadajućim programima i kratkim opisom svakog modula.

Naziv modula	**Programi**	**Kratak opis**
Osnovni ISM modul	ProteinSpektar	Generisanje informacionog spektra proteinske sekvence
Modul za određivanje interaktora	KrosSpektar InteraktorPretraga PikFilterBaze DFTFFTBaza FilterDFTFFTBaze	Nalaženje kandidata iz skupa proteina koji dobro interreaguju sa određenim proteinom
Modul za određivanje interaktivnih domena	AKSkener SetSkener	Određivanje segmenata proteina koji interreaguju sa drugim proteinom
Modul za procenu biološkog efekta mutacija	Mutacije AKSkener LPLPrikaz	Procena uticaja pojedinačnih ili kombinacije mutacija na funkciju proteina

Modul za modifikaciju proteina i dizajniranje peptida	Inverz Kombinator KombCitac	Dizajniranje proteina na osnovu spektralnih karakteristika
Modul za filogenetsku analizu	ISMStablo ISMGraf	Izdvajanje funkcionalnih grupa proteina, detekcija bitnih mutacija i praćenje evolucije funkcije proteina
Modul za pretraživanje molekulskih biblioteka	Chemdb2Alati NiaidPubchemSpoj FormulaKalkulator ValencPotencKalk PubchemParser PubchemTxtParser QSARParser Raspodela Raspodela2D	Pretraživanje molekulskih biblioteka ChemDB i Pubchem, računanje EIIP vrednosti molekula i prikazivanje osnovne statistike
Modul za obradu zapisa sekvenci	SekEditor SekuFasta ProteinBazaSec DNKuProtein FastaMutGen FastaFilter GenomNetFilter	Ručna i automatska obrada tekstualnog zapisa sekvenci, pretraživanje iz baza sekvenci

Pri korišćenju EIIP/ISM platforme za *in silico* dizajniranje lekova postoji određeni tok korišćenja modula koji je prikazan shematski na slici 4.2.2.



Slika 4.2.2. Shematski prikaz toka korišćenja ISM platforme.

Analiza interakcija bioloških molekula i proces pronalaženja lekova zasnovanih na EIIP/ISM platformi, u koje su uključeni principi dalekodosežnih i kratkodosežnih molekulskih interakcija, se sastoji iz dva glavna dela:

1. Dizajniranje peptida i/ili selektovanje malih molekula korišćenjem EIIP/ISM platforme zasnovane na principu dalekodosežnih interakcija. Ceo proces se sastoji od dva nezavisna potprocesa:
   - Dizajniranje peptida:
     1. Programima iz modula za obradu zapisa sekvenci se iz baze proteinskih sekvenci selektuje određeni skup proteina koji učestvuju u određenim biološkim procesima, zatim se sredi, proveri i formatira njihov zapis.
     2. U osnovnom ISM modulu se izračuna informacioni spektar i detektuju specifične spektralne karakteristike proteina.
     3. Generišu se informacije o efektima mutacija i interaktivnim regionima proteina nekim od sledećih postupaka:
        a) Modulom za određivanje interaktora pronađu se povoljni kandidati za interaktore u bazi proteina i modulom za određivanje interaktivnih domena detektuju se regioni proteina koji učestvuju u interakciji.
        b) Iz skupa mutacija identifikuju se bitne mutacije koje menjaju intenzitet interakcije uz pomoć modula za procenu biološkog efekta mutacija.
        c) Modulom za filogenetsku analizu detektuju se bitne funkcionalne grupe proteina i njihove karakteristične mutacije.
     4. Korišćenjem proteinske banke podataka i modula za modifikaciju i dizajniranje peptida, dizajniraju se novi peptidi kao kandidati za lek.
   - Pretraživanje novih kandidata iz baze malih molekula:
     1. Formiraju se test grupe proteina od poznatih lekova koji su testirani i odobreni.
     2. Izračunaju se AQVN i EIIP vrednosti test grupe.
     3. Statističkim metodama se odrede intervali AQVN i EIIP vrednosti, koji su ulazni parametri za sledeći korak.
     4. Iz baze malih molekula se filtriraju i pretražuju novi kandidati pogodni za lek.
2. Standardnim alatima zasnovanim na principu kratkodosežnih interakcija se selektovani mali molekuli i/ili dizajnirani proteini filtriraju i testiraju, na osnovu čega se pronalaze konačni kandidati za lek i terapeutski targeti.

Struktura platforme će u narednim odeljcima biti predstavljena na sledeći način. Svaki modul će biti posebno dat u odeljku sa objašnjenjem svrhe i funkcije modula, kao i spiskom pripadajućih programa. Zatim će se svaki program modula opisati u posebnom pododeljku, gde će se redom navesti: svrha programa, ulazni, izlazni podaci, opis rada programa i algoritam. Pored toga, kod kompleksnijih programa će biti dat opis interfejsa, a za bitnije programe modula će se predstaviti njihova implementacija kroz klasne dijagrame i liste osnovnih klasa.

4.2.3. Osnovni ISM modul

Osnovni ISM modul sadrži program ProteinSpektar. Funkcije modula su generisanje informacionog spektra određene proteinske sekvence primenom FFT ili DFT transformacije i određivanje značajnih pikova spektra, kao i detekcija njihove karakteristične frekvence.

4.2.3.1. Program *ProteinSpektar*

Svrha:

- Generisanje informacionog spektra proteina
- Upoređivanje spektara dobijeni FFT i DFT transformacijama
- Određivanje spektralnih karakteristika proteina, tj. značajnih pikova spektra
- Uvođenje oznake bojom (eng. *Color code*) kao novog načina grafičke prezentacije ISM spektara.

Ulaz:

- Datoteka sekvenci tipa *.seq
- Minimalna dužina sekvence (L)

Izlaz:

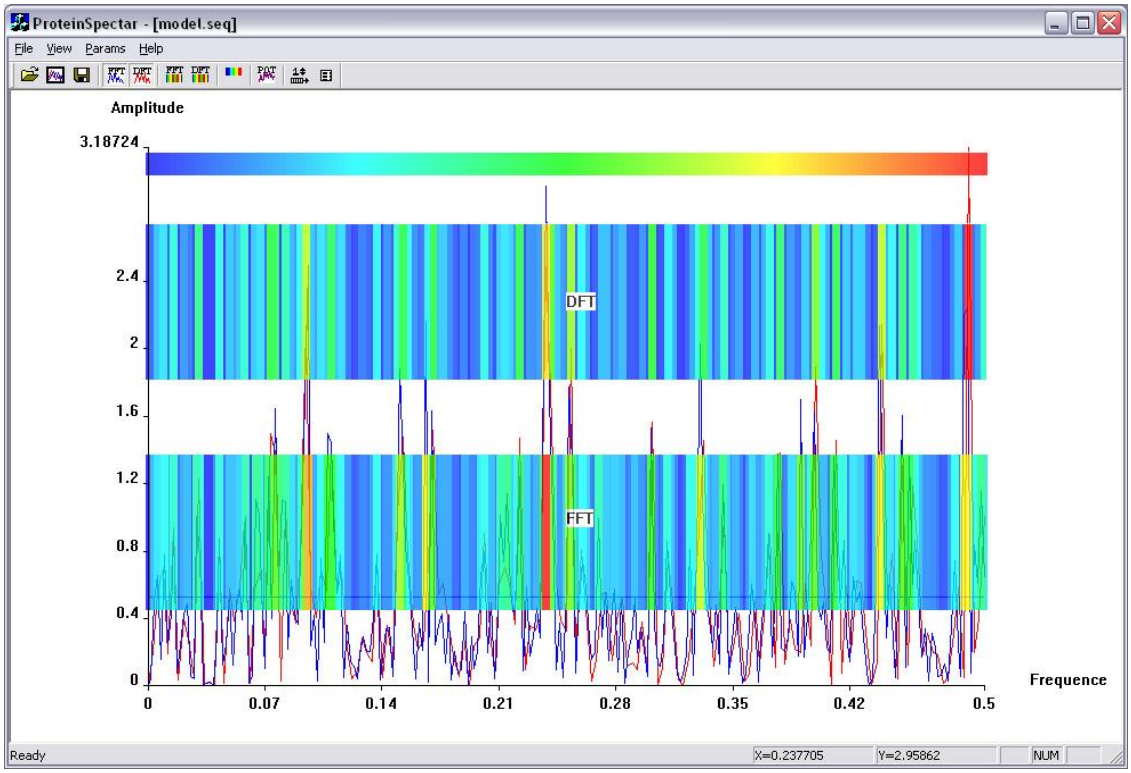
- Slika trenutno prikazanih grafika spektara u datoteci tipa *.bmp

- Tabele DFT i FFT spektara sa poljima: amplituda, signal/šum, pik i niz EIIP signala sa vrednostima EIIP potencijala, koji su u tekstualnom formatu ili u formatu za program *Origin*.

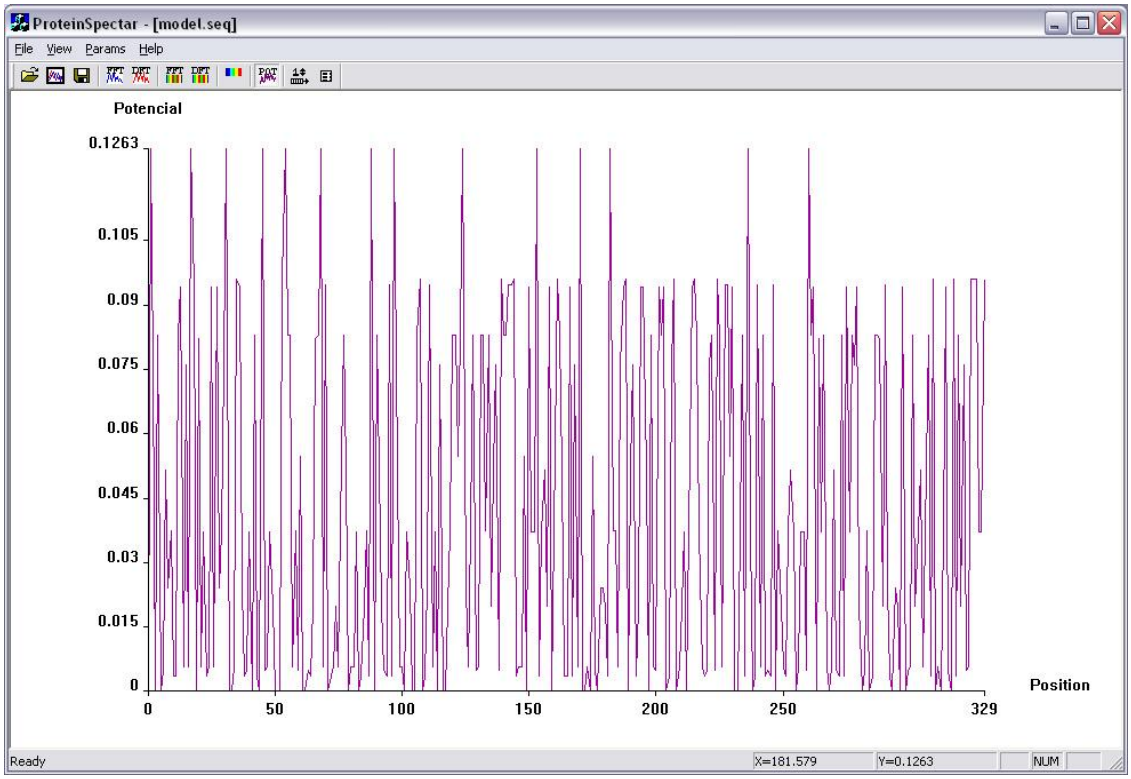
Za ulaznu proteinsku sekvencu program računa ISM spektar koristeći FFT i DFT transformacije. Ako je zadat parametar *L*, prvo se signal produži do dužine *L*, a u slučaju FFT transformacije se signal produži do najbližeg stepena dvojke. Na grafiku se mogu istovremeno iscrtati FFT spektar plavom bojom, kao i DFT spektar koji je crvene boje. Može se iscrtati i grafik EIIP signala zadate sekvence. Grafici se mogu svi zajedno iscrtati u prozoru. U desnom delu statusne trake ispisuju se koordinate grafika na poziciji miša iznad grafika. Tabele celih FFT i DFT spektara, vrednosti pikova sortirani prema amplitudama i vrednosti EIIP signala, mogu se snimiti u tekstualnu datoteku ili datoteku u formatu za program *Origin*.

Oznaka bojom za informacioni spektar

Kao posebna opcija programa uvedeno je novo prikazivanja spektara u obliku **oznake bojom (eng. *Color code*).** Označavanje informacionog spektra bojom predstavlja se horizontalnom trakom sa vertikalnim poljima čija Y pozicija odgovara vrednosti frekvence, a boja polja predstavlja vrednost amplitude spektra na toj frekvenci. Boje spektra su različite u zavisnosti od izbora skale. Prednost oznake bojom je manji prostor potreban za prikaz, pregledniji prikaz spektra za čoveka, kao i lakše upoređivanje više spektara.



Slika 4.2.3. Prozor programa *ProteinSpektar* sa prikazanim DFT i FFT informacionim spektrima i oznakama bojom.



Slika 4.2.4. Prozor programa *ProteinSpektar* sa prikazanim EIIP signalom.

4.2.4. Modul za određivanje interaktora

Modul za određivanje interaktora sadrži programe: KrosSpektar, InteraktorPretraga, PikFilterBaze, DFTFFTBaza i FilterDFTFFTBaze. Modul se primenjuje za određivanje interaktora (kandidata, iz skupa proteina, koji dobro interreaguju sa zadatim proteinom), kao i za procenu interakcija.

4.2.4.1. Program *KrosSpektar*

Svrha:

- Traženje zajedničkih bioloških karakteristika grupe proteina, traženjem pikova u kros-spektru.
- Pretraga baze proteina za interaktorima koji imaju određenu zajedničku osobinu sa drugom grupom proteina.
- Analiza grupe sličnih proteina, njihovo upoređivanje, procena zastupljenosti određene karakteristike svake sekvence posebno i grupno.

Ulaz:

- Baze proteinskih sekvenci koje mogu biti oblika:
    - Proteinska sekvenca u SEQ formatu
    - Baza proteina u FASTA formatu
    - Baza proteina u SwissProt formatu
    - Sve *.seq datoteke iz direktorijuma
    - Direktan unos proteinske sekvence
- Snimljen kros-spektar programom *KrosSpectar* iz datoteke tipa *.cs
- Parametri za generisanje krosspektra:
    - *Deltafrek* interval - u kojem se intervalu oko određene frekvence uzimaju u obzir i susedne frekvence.
    - *Metoda* - način fiksiranja spektara i množenja amplituda
    - Normiranje amplituda na maksimalnu vrednost jedan.

Izlaz:

- Ceo krosspektar dokumenta se može snimiti u datoteku tipa *.cs, koja sadrži niz svih sekvenci sa opisima i parametrima za svaki krosspektar.
- Ceo spektar u formatu tabele (polja razdeljena tab simbolima) sa vrednostima: *Amplituda, S/N, Šum*.
- Pikovi spektra sa vrednostima: *amplituda, S/N, Šum, broj pika*
- Slika grafika selektovanog spektra ili kros-spektra
- U *KrosSvi* modulu: posebna tabela isfiltriranih sekvenci sa vrednostima amplituda kros-spektara i rednim brojem pika u slučaju uključenog filtera za pikove.

Program računa kros-spektre više proteina, produžujući EIIP signale na dužinu do najkraćeg stepena dvojke zbog FFT algoritma i uzimajući u obzir parametre za generisanje krosspektra. Postoji poseban mod programa *PunKros* koji ima mogućnost da izračuna ceo kros-spektar za neograničen broj sekvenci iz ulazne baze. Za svaki spektar i kros-spektar prikazuju se podaci spektra sa prvih deset pikova, grafik, parametri u slučaju kros-spektra i sekvenca u slučaju spektra.

Struktura za predstavljanje kros-spektara

Pošto se kros-spektar sastoji od običnih spektara, ali takođe može biti sastavljen od drugih kros-spektara jer je kros-spektar i sam spektar koji nema svoju sekvencu, struktura kros-spektra je predstavljena hijerarhijski u drvolikom obliku. Sekvence se mogu dodavati u kros-spektar, i takođe posebno brisati iz njega. Pri svakom slučaju se kros-spektar preračuna. Postoji više načina dodavanja sekvenci u dokument, tj. kros-spektar: iz baza proteina, direktnim unosom, jedna po jedna, ili više istovremeno. Ako se dodaje više sekvenci, postoji mogućnost normiranja amplituda da ne bi došlo do prekoračenja pri računanju, tj. množenju amplituda u slučaju prevelikog broja sekvenci. Pri selektovanju svakog kros-spektra posebno se mogu videti podaci, parametri i grafik za svaki spektar i kros-spektar.

KrosSvi opcija programa *KrosSpektar*

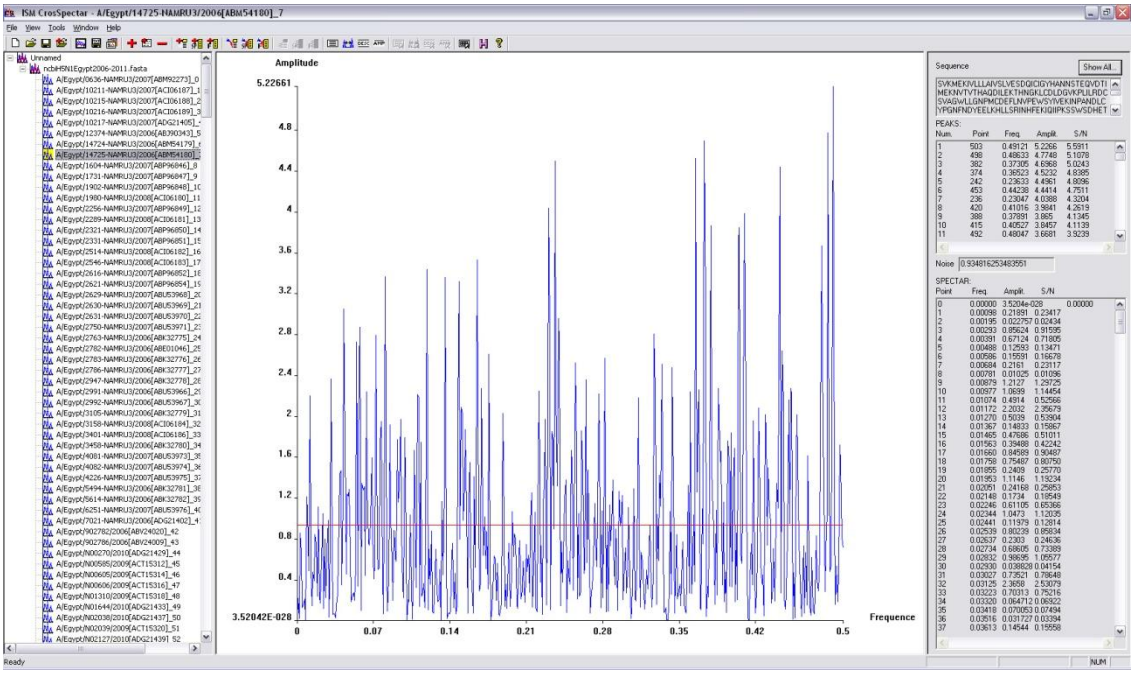
Postoji posebna opcija programa, gde se mogu izračunati kros-spektri trenutnog spektra iz dokumenta redom sa svim proteinima iz druge ulazne baze, uz moguće filtriranje po piku na zadatoj frekvenci krosspektra i rezultat prikazati u obliku tabele ili snimiti u datoteku (slika 4.2.6). Ova opcija omogućava pretraživanje (skrining) proteina iz baze koji u konsenzus informacionom spektru sa zadatim proteinom imaju visoku amplitudu na traženoj frekvenci.

Osim 20 standardnih kiselina, slovo 'X' označava nepoznatu kiselinu, koja se u procesu kodiranja EIIP vrednostima kodira u vrednost nula u spuštenom EIIP signalu za srednju vrednost. U opisu kros-spektra se ispisuju informacije o nazivu ulazne baze i vrednosti parametara krosspektra, kao i broj sekvenci u kros-spektru.

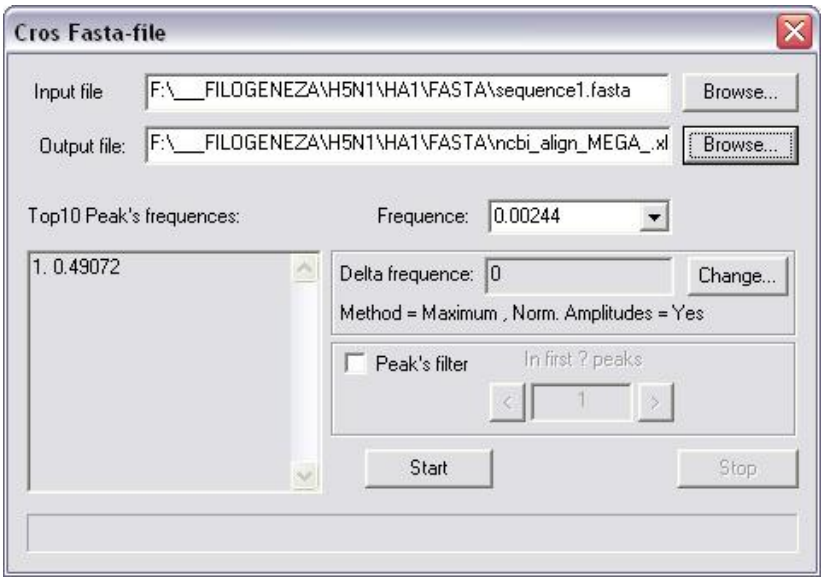
Parametri za izračunavanje kros-spektra

Pri računanju amplitude kros-spektra zbog finijeg podešavanja i u zavisnosti od cilja analize, množenje amplituda uključenih spektara se može izvesti na nekoliko načina. Zbog toga su uvedeni sledeći parametri: (i) prvi parametar je interval *deltafrek* koji određuje u kojoj okolini oko frekvence će se, osim amplitude na toj frekvenci spektra, uzimati amplitude susednih frekvenci kao činioci u množenju pri računanju amplitude kros-spektra, (ii) drugi parametar je *metoda* koji određuje način množenja, a koji može biti:

- Maksimum - maksimalna amplituda svih u frekventnom intervalu
- Srednja - srednja vrednost amplituda svih frekvenci u intervalu
- Max & fiksirana prva - za prvu sekvencu se gleda samo centralna amplituda, tj. njen interval je nula i tada se fiksira ta sekvenca kao bitna, a za ostale se računa kao za metodu *Maksimum*
- Srednja & fiksirana prva - za prvu sekvencu se gleda kao kod *Max & fiksirana prva*, a za ostale kao kod metode *Srednja*.



Slika 4.2.5. Osnovni prozor programa *KrosSpektar* sa panelima: levi panel predstavlja listu sekvenci, na srednjem panelu je slika spektra označene sekvence, desno su numeričke vrednosti spektra i sortirani pikovi.



Slika 4.2.6. Prozor dijaloga opcije *KrosSvi* programa *KrosSpektar*.

Algoritam za generisanje ISM kros-spektra:

Ulaz: skup proteinskih sekvenci.

Izlaz: ISM kros-spektar

1. Utvrđivanje dužine *Dmax* najduže sekvence i vrednosti *L* najmanjeg stepena dvojke koji je veći ili jednak *Dmax*.
2. Za svaku sekvencu S_i se generiše njen ISM spektar X_i pomoću sledećih koraka:
   a. Kodiranje sekvence S_i u signal P_i preko EIIP vrednosti aminokiselina.
   b. Spuštanje signala *Pi* za srednju vrednost, tako da nov signal ima srednju vrednost nula.
   c. Produžavanje nulama signala sa desne strane do dužine *L* čime se svi signali dovode na istu dužinu i svi spektri u istu rezoluciju.
   d. Prevođenje signala u vektor kompleksnih brojeva tako što se signal dodeli realnom delu a imaginarni postavi na nulu.
   e. Brza Furijeova transformacija (FFT) signala P_i u spektar X_i. Uzima se samo prva polovina spektra jer je simetričan. Interval frekvenci se skalira na interval [0,0.5].
   f. Računa se vektor amplituda A_i spektra X_i : $A_i[j] = |S_i[j]|$, $j = 0 .. L/2–1$, $i = 1 .. N$, gde je N broj sekvenci, $S_i[j]$ kompleksna vrednost spektra *Xi* na *j* koordinati.
3. Izračuna se vektor kros-spektra *CS* svih spektara: $CS[j] = A_0[j] * ..* A_N[j]$, $j = 0 .. L/2–1$, gde je N broj svih ulaznih sekvenci.

Korisnički interfejs programa *KrosSpektar*

Program se sastoji iz dva osnovna moda:

- Glavni mod
- PunKros mod

Modovi se mogu birati samo na početku kada nije učitan nijedan dokument, odnosno kada je dokument prazan (*File→New*).

Glavni prozor se sastoji iz tri panela (slika 4.2.5):

- Levi panel (*CSTree*) sadrži listu svih spektara (jednostavnih i podkrosspektara) oraganizovanih u drvenoliku strukturu. Postoje dve vrste spektara: jednostavni spektar i krosspektar koji se sastoji od spektara. Ova drvenolika struktura predstavlja kompozitni šablon (eng. *composite pattern*) spektara, i kao takva je implementirana u izvornom kodu ISM paketa.
- Srednji panel (*GraphView*) služi za iscrtavanje grafika spektra selektovane sekvence. Svaki spektar se izračunava pri prvom selektovanju. Status računanja je prikazan u levom status baru. Pri prelazu mišem preko grafika, u desnoj statusnoj traci se ispisuju koordinate grafika na poziciji miša. Klikom miša na pik u grafiku, selektuje se pik spektra i ispisuju se vrednost frekvence, amplituda i redni broj selektovanog pika.
- U desnom panelu se ispisuju rezultati selektovanog spektra: pikovi i vrednosti kros-spektara i u zavisnosti od selektovanog spektra:
    - Parametri krosspektra, ako je selektovani spektar kros-spektar
    - Proteinska sekvenca koja odgovara tom spektru, ako je selektovan jednostavan spektar.

PunKros* mod (eng. *FullCross*) programa *KrosSpektar

Ovo je specijalan mod za izračunavanje kros-spektra velikog broja sekvenci bez ograničenja. Ulazni format sekvenci može biti: *Folder* / *SwissProt* / *Fasta*, gde se u slučaju *Folder* uzimaju sve datoteke tipa **.seq* iz zadatog foldera. Opcija *Tools→CrossAll Folder/Prot/Fasta* otvara *CrossAll* dijalog u kojem se mogu podesiti parametri kros-spektra: delta-interval frekvenci, metoda i normalizacija amplituda koja je podrazumevana, jer može doći do prekoračenja. Nakon završetka izračunavanja kros-spektra, iscrtava se grafik krosspektra i jedine dozvoljene opcije na meniju su snimanje i eksportovanje podataka i grafika.

Glavni mod programa *KrosSpektar*

File meni programa *KrosSpektar*

Akcije menija omogućavaju otvaranje krosspektar dokumenta i snimanje trenutnog krosspektra *RCS* u dokumentu (eng. *RootCrossSpectrum - RCS*). Postoji više načina za dodavanje novih sekvenci u trenutni dokument:

- Jedna sekvenca iz *.seq datoteke
- Direktno ukucavanjem zapisa jedne sekvence
- Sekvence iz svih *.seq datoteka zadatog direktorijuma
- Sve sekvence iz datoteke u SwissProt formatu
- Sve sekvence iz datoteke u FASTA formatu

U slučaju dodavanja više datoteka, može se izabrati opcija normalizovanje amplituda. Pri preračunavanju kros-spektara u levoj statusnoj traci se prikazuje status računanja.

Akcija *Briši trenutni spektar* briše izabrani spektar iz dokumenta i preračunava kros-spektre.

Akcija *Eksportovanje* omogućava snimanje slike trenutno izabranog spektra, kao i vrednosti pikova i spektra u formatu tabele (posebno za program *Origin*).

Tools meni programa *KrosSpektar*

Opcije *CrossFolder/Prot/Fasta* se mogu koristiti kada dokument nije prazan, tj. kada sadrži već neki spektar (RCS). U *CrosFolder/Prot/Fasta* dijalog se unose parametri: selektovana frekvenca *F*, parametri kros-spektra, ulazni direktorijum/datoteka, filter *N* prvih pikova. Za svaku ulaznu *S*, iz ulazne baze, računa se kros-spektar *CS* koji se sastoji od spektra dokumenta *RCS* i ulazne sekvence *S*. Ako je zadat filter, krosspektar se filtrira prema uslovu: da li je vrednost na selektovanoj frekvenci *F* u kros-spektru *CS* među prvih *N* pikova. Ako nije zadat filter, sekvenca svakako ulazi u rezultat. Rezultat je tabela koju čine vrednosti krosspektra za svaku isfiltriranu sekvencu sa kolonama: opis, sekvenca, amplituda (na frekvenci *F*), signal/šum, šum, broj pika.

Window* meni programa *KrosSpektar

Stavka menija *RCS Sekvence* prikazuje nazive i opise jednostavnih spektara koji se nalaze u RCS. Dijalog *RCS Parametri* omogućava izmenu parametara za RCS: frekventni interval, metoda, normalizacija amplituda. Parametri se mogu primeniti samo na RCS, ili na sve podkrosspektre. Stavka *RCS Opis* otvara dijalog za prikaz i izmenu opisa za RCS. Stavka *RCS Amplitudes Average* prikazuje srednje vrednosti amplituda svih jednostavnih spektara u RCS. Stavka *SCS Sequences* prikazuje nazive i opise sekvenci u trenutno selektovanom spektru SCS (eng. *currently Selected CrossSpectra - SCS*). U dijalogu *SCS Parameters* se mogu izmeniti parametri za SCS: frekventni interval, metoda, normalizacija amplituda. Stavka *SCS Amplitudes Average* prikazuje srednje vrednosti amplituda svih jednostavnih spektara u SCS. Stavka *Selected Sequence* prikazuje opis i sekvencu selektovanog jednostavnog spektra.

4.2.4.2. Program *InteraktorPretraga*

Svrha:

- Pretraživanje potencijalnih interaktora iz velike proteinske baze, koji dobro interreaguju sa svim proteinima iz male baze sa sličnom funkcijom.

Ulaz:

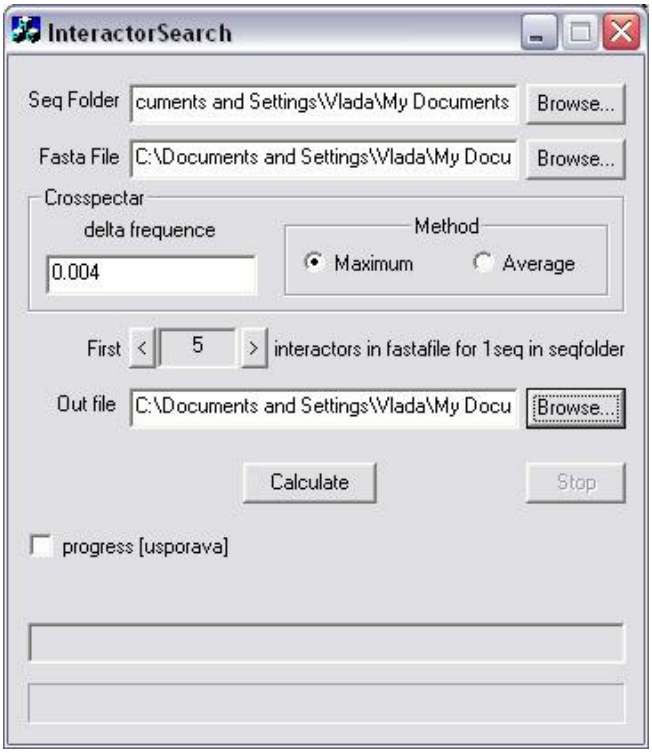
- Sve datoteke tipa SEQ iz direktorijuma (mala baza)
- Datoteka proteina u FASTA formatu (velika baza)
- Parametri za generisanje kros-spektra
    - *Deltafrek interval*: određuje u kojem intervalu se oko određene frekvence uzimaju u obzir i susedne frekvence.
    - *Metoda*: određuje način fiksiranja spektara i množenja amplituda
- Broj prvih *N* interaktora iz FASTA datoteke za jednu sekvencu iz *.seq datoteke

Izlaz:

- Tabela interaktora iz druge grupe sekvenci, koja se sastoji iz dva dela:
    - Za svaku sekvencu iz prvog skupa (sa poljima: *ime_fajla*, *opis*), piše prvih *N* interaktora sa poljima vrednosti krosspektra: *S/N*, *frekvenca*, *fasta_opis*, a zatim ispisuje sortirane interaktore po broju ponavljanja sa poljima: broj_*ponavljanja*, *fasta_opis*.
    - Sortirani svi interaktori iz drugog ulaznog skupa po broju ponavljanja sa poljima: ∑*(S/N)*, *fasta_opis*.

Za dva ulazna skupa proteinskih sekvenci, program traži iz druge grupe (veća baza) najbolje interaktore sa svim sekvencama prve grupe (manja baza), tako što uparuje redom svaku sekvencu iz prvog skupa sa svakom sekvencom iz drugog skupa i računa njihov krosspektar. Krosspektar se izračunava na osnovu zadatih parametara kao kod programa *KrosSpektar*. Posmatraju se najveće vrednosti amplituda i signal/šum (eng. *Signal/Noise*, *S/N*) na prvom piku u krosspektrima. Na osnovu njih se vrši sortiranje interaktora pomoću dve slične strategije:

1. Za svaku sekvencu iz direktorijuma (prvog skupa) se sortira samo prvih *N* interaktora iz drugog skupa sekvenci prema S/N na prvom piku njihovog kros-spektra. Zatim se za svaki interaktor iz druge grupe prebroji sa koliko sekvenci iz prve grupe je u skupu od prvih *N* i interaktori se na kraju sortiraju prema tom broju ponavljanja.
2. Za svaku sekvencu interaktora iz drugog skupa se sumira vrednost signal/šum prvog pika kros-spektara u kojima učestvuje i interaktori se sortiraju prema tim sumama.



Slika 4.2.7. Prozor programa *InteraktorPretraga*.

Algoritam:

Ulaz: skup proteinskih sekvenci *G1*, skup sekvenci *G2* kao mogućih interaktora sa *G1*, parametri za generisanje krosspektra

Izlaz: tabela sortiranih interaktora iz *G2* prema interaktivnom potencijalu

1. Za svaku sekvencu *I* iz *G2*:
   a. *SN(I)* = 0, gde je *SN(I)* suma signal/šum na prvom piku krosspektra *CS*.
   b. Za svaku sekvencu *S* iz *G1*:
      i. Generisati krosspektar *CS* od sekvenci *S* i *T*.
      ii. *SN(I)* = *SN(I)* + *Sg(1)*, gde je *Sg(1)* vrednost signal/šum na prvom piku krosspektra *CS*.
2. Sortirati interaktore *I* prema vrednostima *SN(I)*.

4.2.4.3. Program *PikFilterBaze*

Svrha:

- Pretraživanje proteina iz baze sekvenci po određenoj frekvenci koja karakteriše neku biološku funkciju za identifikaciju potencijalnih proteina sa tom biološkom funkcijom.

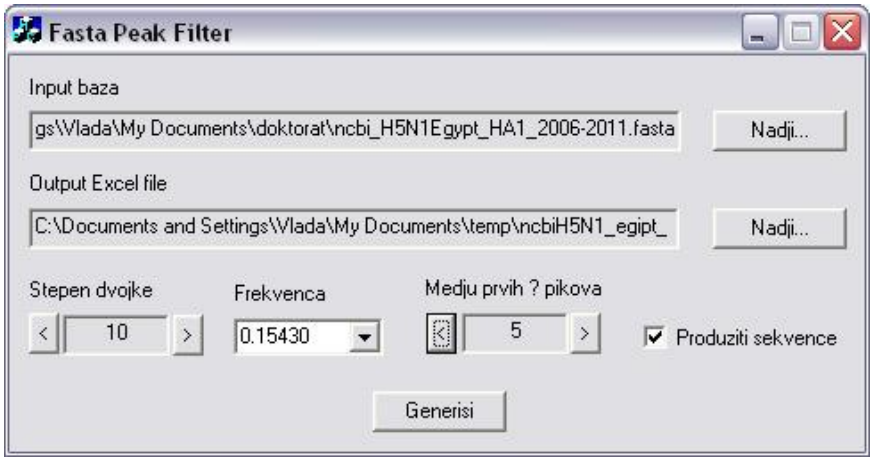
Ulaz:

- Datoteka proteina u FASTA ili SWISSPROT formatu ili sve datoteke tipa *.seq iz zadatog direktorijuma
- Ulazni parametri
    - Frekvenca (*F*)
    - Broj pikova (*N*)
    - Opcija produžavanja sekvence do dužine *L*

Izlaz:

- Tabela proteina koji su zadovoljili uslov filtriranja. Svaki red tabele odgovara jednoj sekvenci. Kolone su: ID, frekvenca (najbliža), amplituda, S/N, broj pika, opis, sekvenca.

Za svaki protein iz ulazne baze se izračuna njegov ISM spektar, uz prethodno poduživanje signala do zadate dužine *L* ako je data, pronalazi se amplituda na frekvenci *najbližoj* zadatoj frekvenci *F*. Ako je amplituda među prvih *N* zadatih pikova spektra, sekvenca zadovoljava uslov filtriranja. Traži se *najbliža* frekvenca jer rezolucije svih spektara mogu biti različite dužine, osim ako je zadato *L* u kom slučaju su svi spektri iste rezolucije.



Slika 4.2.8. Prozor programa *PikFilterBaze*.

4.2.4.4. Program *DFTFFTBaza*

Svrha:

- Računanje ISM spektara određene grupe proteina i snimanje samo prvih *N* pikova svakog spektra za kasnije brže pretraživanje i filtriranje baze uz pomoć programa *FilterDFTFFTBaze.*

Ulaz:

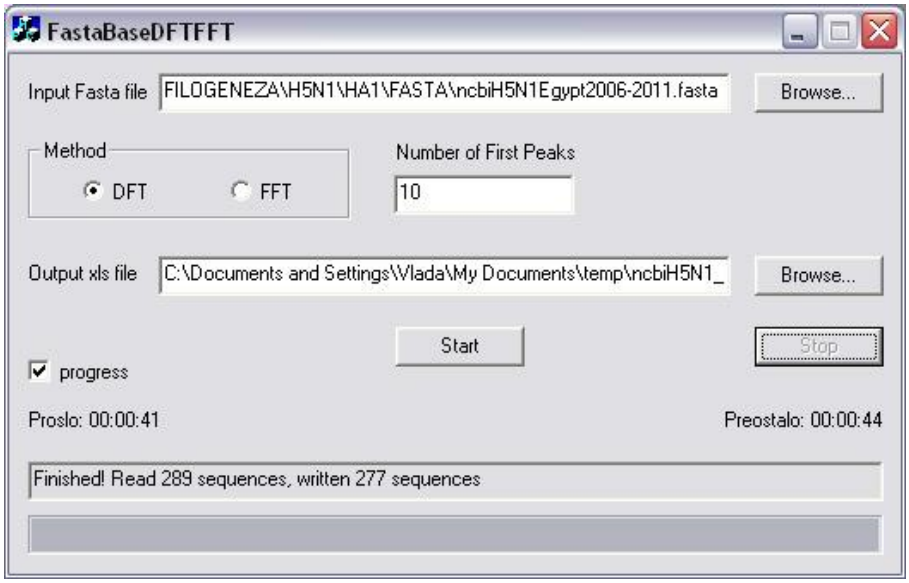
- Baza sekvenci u FASTA formatu
- Transformacija: DFT ili FFT
- Broj prvih pikova (*N*)

Izlaz:

- Tabelarni format (polja razdvojena tab simbolom) gde svaki red odgovara jednoj sekvenci iz ulazne baze. Kolone su: ID, opis, $frek_1$, amp_1, s/n_1, ... $frek_N$, amp_N, s/n_N, šum, sekvenca. *N* je broj pikova. Tabela se može sačuvati u tekstualnoj datoteci.

Za svaku sekvencu iz ulazne baze, program računa DFT ili FFT spektar, u zavisnosti od izbora metoda, i piše u izlaznu bazu prvih *N* pikova (gde je *N* zadat broj prvih pikova), sortiranih redom od najvišeg pika. Za svaki pik ispisuje njegovu frekvencu, amplitudu, signal/šum i na kraju ispiše šum celog spektra. Kada se jednom

obradi baza, kasnije se može filtrirati po intervalu frekvence prvog pika ili prvih *K* pikova u *Excel* programu ili programu *FilterDFTFFTBaze*.



Slika 4.2.9. Prozor programa *DFTFFTBaza*.

Algoritam:

Ulaz: baza u FASTA formatu, metoda *M*, broj prvih pikova *N*.

Izlaz: tabela sekvenci sa vrednostima amplituda i S/N na prvih *N* pikova dobijeni metodom FFT/DFT.

1. Za svaku sekvencu *S* iz baze:
    1) Izračunati EIIP signal *P* sekvence *S*
    2) Izračunati ISM spektar *W* signala *P* pomoću FFT ili DFT transformacije, u zavisnosti od zadate metode *M*. U slučaju FFT, signal se produži do najbližeg stepena dvojke.
    3) Ispisati: frekvencu, amplitudu i signal/šum za svaki pik od prvih *N* pikova iz spektra *W*, i šum celog spektra *W*.

4.2.4.5. Program *FilterDFTFFTBaze*

Svrha:

- Brzo pretraživanje baze sekvenci sa izračunatim spektrima (pikovima), po određenoj frekvenci koja karakteriše neku biološku funkciju, za identifikaciju potencijalnih proteina koji poseduju tu funkciju.

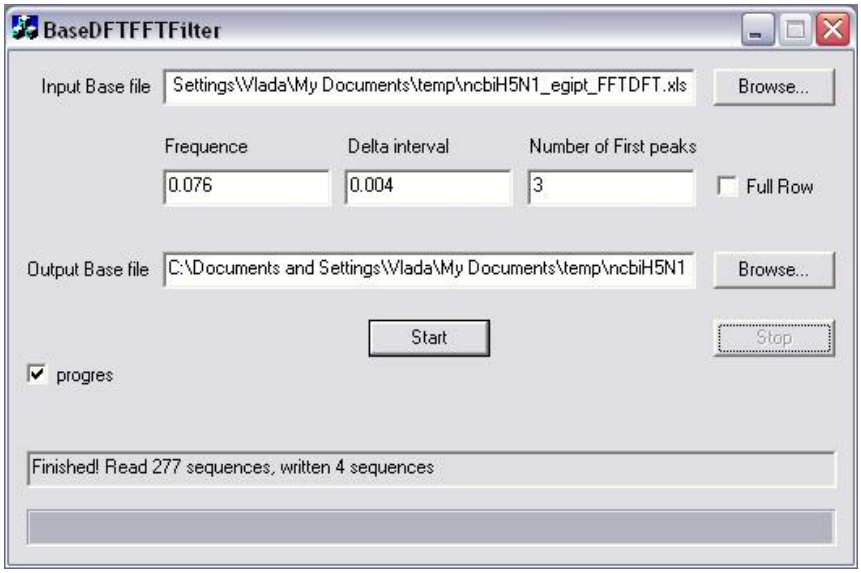
Ulaz:

- Tabela koja je rezultat programa *DFTFFTBaza*
- Frekvenca spektra po kojoj se filtrira (*frekvenca*)
- Delta interval frekvence oko koje se pretražuje (*delta_interval*)
- *Broj pikova* - među koliko prvih pikova da pretražuje

Izlaz:

- Tabelarni format (polja razdvojena tab simbolom) gde je prvi red zaglavlje, tj. prvi red ulazne tabele. Svaki sledeći red je sekvenca koja zadovoljava kriterijum: ako je zadat parametar *Pun red*, prepisuje se ceo red iz ulazne tabele (tj. svi pikovi sa svojom amplitudom i S/N), ako nije zadat, onda se ispisuje u formatu sa kolonama: *ID, Opis, RedniBrojPika, Frekvenca, Amplituda, Signal/Šum, Šum, Sekvenca*, odnosno samo onaj pik (sa svojom amplitudom i S/N) koji je u zadatom intervalu frekvence.

Program obrađuje, tj. filtrira datoteku u formatu tabele (polja razdvojena tab simbolom) koja je rezultat programa *DFTFFTBaza*. Za zadatu frekvencu i *delta_interval*, program filtrira one sekvence koje među prvih *N* zadatih pikova u tabeli, imaju pik na frekvenci koja je u intervalu [*frekvenca – delta_interval*, *frekvenca + delta_interval*].



Slika 4.2.10. Prozor programa *FilterDTFFTBaze*.

Algoritam:

Ulaz: tabela koja je rezultat programa *DFTFFTBaza*, frekvenca *F*, delta-interval *d*, *N* broj prvih pikova, parametar *prvi_red*.

Izlaz: tabela sekvenci sa vrednostima amplituda i S/N.

1. Za svaku sekvencu *S* iz tabele:
   a. Ispitati da neki od prvih *N* pikova ima frekvencu u intervalu [*F* - *d*, *F* + *d*]. Ako ima, onda ako je zadat *prvi_red* prepisati ceo red, inače samo onaj pik koji to zadovoljava i njegov redni broj.

4.2.5. Modul za određivanje interaktivnih domena

Modul za određivanje interaktivnih domena se koristi za određivanje segmenata proteina koji učestvuju u interakciji sa drugim proteinom. Modul sadrži programe AKSkener i SetSkener.

4.2.5.1. Program *AKSkener*

Svrha:

- Identifikacija funkcionalnih delova proteina, tj. onih segmenata sekvence koji su zaduženi za određenu biološku funkciju koja se ispoljava zadatom frekvencom.

Ulaz:

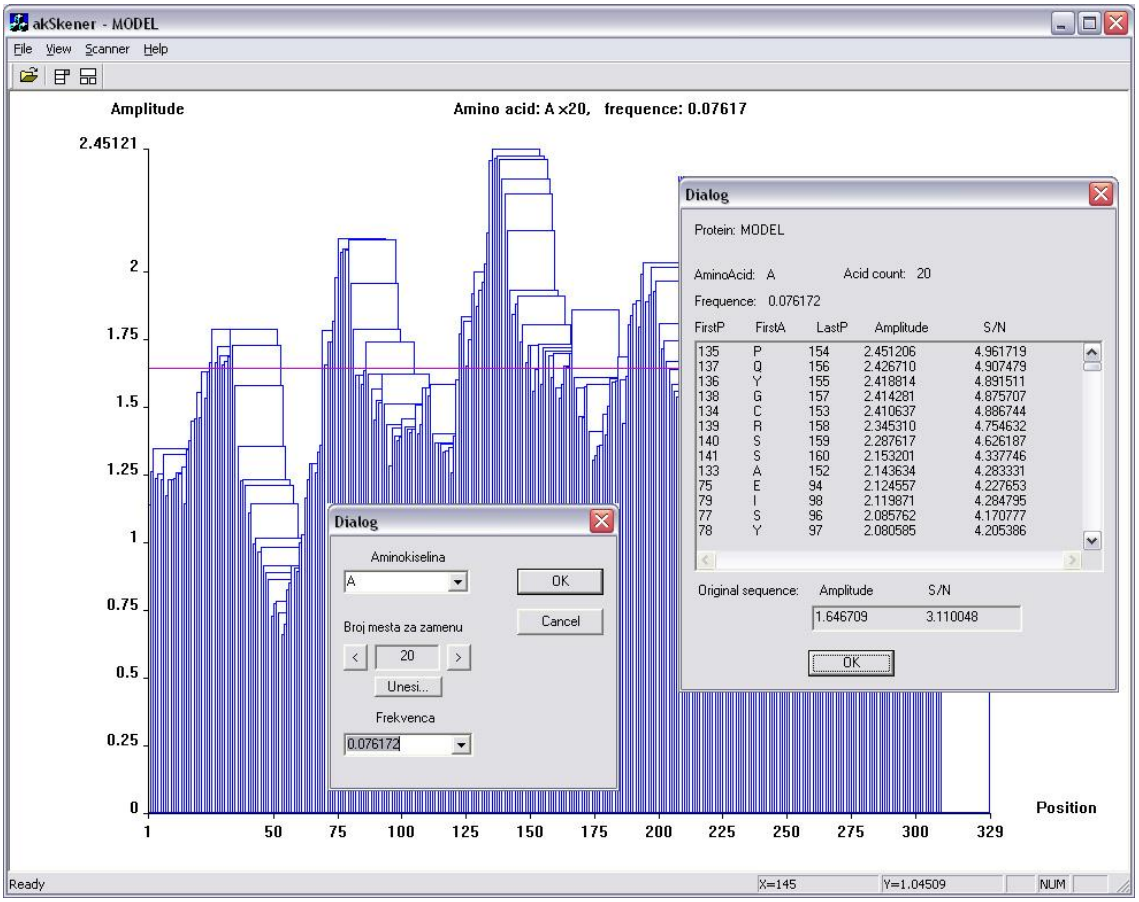
- Proteinska sekvenca u SEQ formatu
- Frekvenca iz opsega rezolucije ISM spektra zadatog proteina
- Dužina prozora regiona koji se transformiše

Izlaz:

- Grafik koji predstavlja raspodelu amplituda na zadatoj frekvenci spektara svih izmenjenih sekvenci. Amplitude su predstavljene pravougaonicima gde visina odgovara vrednosti amplitude, a levi i desni krajevi odgovaraju pozicijama početka i kraja izmenjenog regiona u originalnoj sekvenci.
- Tabela sortirana u opadajućem redosledu po amplitudama, sa poljima: početna, krajnja pozicija izmenjenog regiona, vrednost amplitude, signal/šum, prva kiselina originalne sekvence na početnoj poziciji regiona.

Program prvo definiše rezoluciju svih mogućih frekvenci ulaznog proteina tako što mu produži sekvencu na najkraći stepen dvojke zbog FFT transformacije. Iz tog skupa frekvenci se bira jedna za koju će program računati amplitude. Originalna sekvenca se transformiše tako što se na *N* uzastopnih pozicija unesu zamene sa zadatom aminokiselinom, gde je *N* zadata dužina prozora. Iterativno se generiše skup novih sekvenci tako što se za početak pozicije od koje se počinje zamena uzima redom od prve do poslednje moguće pozicije u ulaznoj sekvenci. Za zamenjenu kiselinu može se izabrati bilo koja od dvadeset kiselina, kao i specijalni simbol: prazno mesto „-“, kada se *N* uzastopnih pozicije odsecaju iz proteina, ili stop-kod „~“, u kom slučaju se sve iza početne pozicije za zamenu odseca iz proteina. U oba slučaja se nova sekvenca produži do dužine originalne kako bi se očuvala rezolucija. Za dužinu prozora se može izabrati vrednost 1, i tada se redom jedna po jedna kiselinska pozicija menja sa zadatom

kiselinom. Za svaku novodobijenu sekvencu se izračuna ISM spektar i vrednost amplitude i odnosa signal/šum na zadatom frekvenci.



Slika 4.2.11. Prozor programa *AKSkener*.

Algoritam:

Ulaz: Sekvenca *S*, frekvenca *F*, dužina prozora *d*, aminokiselina za zamenu *K*.

Izlaz: grafik amplituda, tabela sekvenci i pozicija izmenjenih regiona sortirana po amplitudama.

1. Proširiti dužinu originalne sekvence *S* na najkraći stepen dvojke veći od dužine sekvence i definisati skup svih mogućih frekvenci ISM spektra tog proteina.
2. Za svaku aminokiselinsku poziciju *X* u originalnom proteinu *S*, gde *X* ide od 1 do *L-d+1*, gde je *L* dužina sekvence *S*:
   a. Generisati novu sekvencu *P* iz originalne, zamenom regiona od pozicije *X* do *X+d* gde je *d* zadata dužina prozora, sa zadatom aminokiselinom *K* gde

se u slučaju K='-' region odseca, a u slučaju K='~' odseca i sve iza regiona.

b. Za novu sekvencu *P* izračunati EIIP signal, zatim ISM spektar *W* pomoću FFT transformacije.
c. Izračunati amplitudu *Ai* i signal/šum na zadatoj frekvenci *F* iz spektra *W*.

3. Sortirati sekvence *P* prema amplitudama *Ai*.

4.2.5.2. Program *SetSkener*

Svrha:

- Identifikacija regiona u proteinu koji imaju najveći uticaj na određenu biološku karakteristiku predstavljenu karakterističnom frekvencom ISM spektra.

Ulaz:

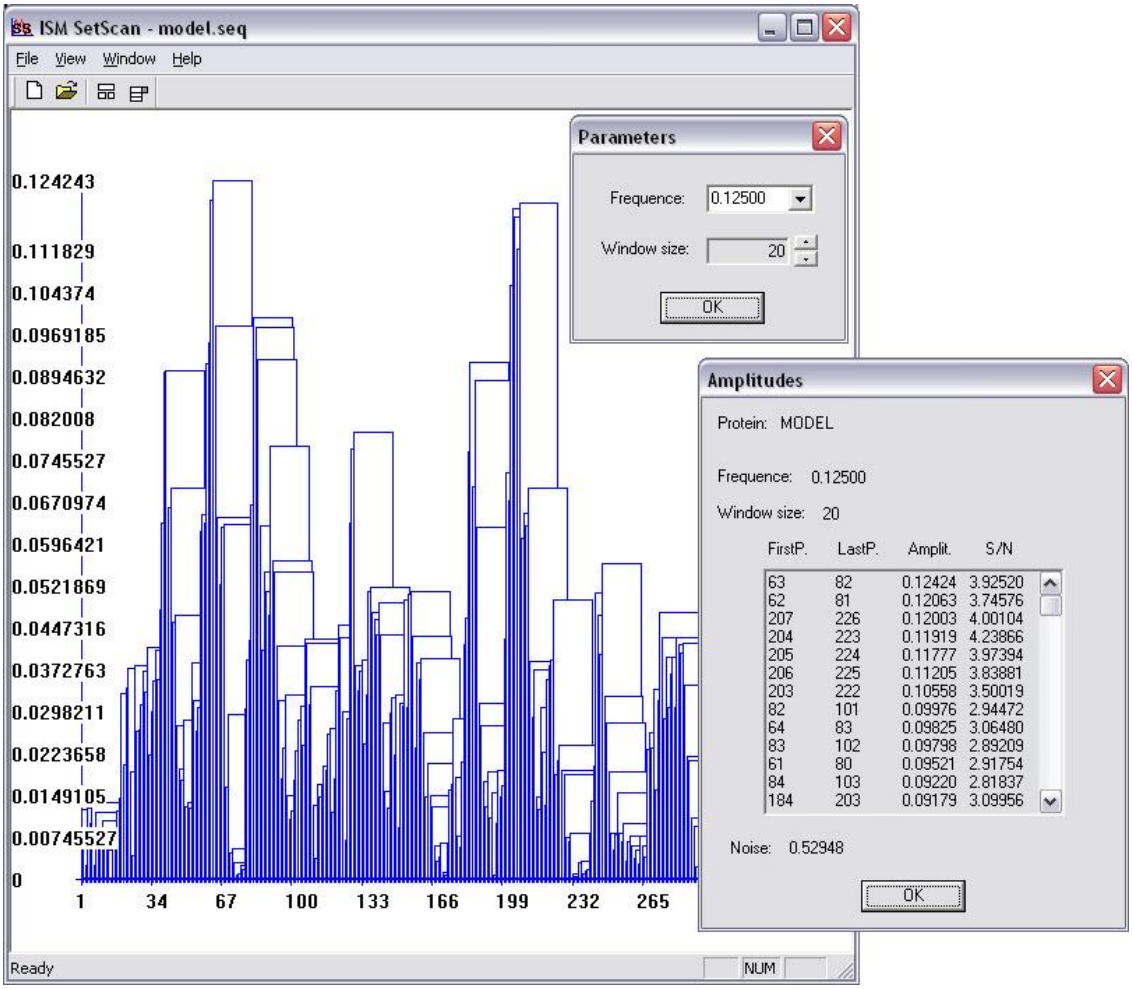
- Datoteka tipa *.seq
- Dužina prozora (podsekvence) (*L*)
- Frekvenca (*F*)

Izlaz:

- Grafik koji predstavlja vrednosti amplituda ISM spektra podsekvenci na frekvenci *F*, po njihovim pozicijama u originalnoj sekvenci. Može se snimiti kao datoteka u *.bmp formatu.
- Tabela podsekvenci sa poljima: prva, poslednja pozicija u sekvenci, amplituda, signal/šum. Moguće je filtrirati samo one podsekvence sa vrednostima signal/šum>1 i/ili sortirati po amplitudama. Tabela se može snimiti u tekstualnu datoteku ili u tabelarni format pogodan za program *Origin*.

Na osnovu izabrane dužine *L*, izabere se karakteristična frekvenca *F* iz skupa mogućih frekvenci ISM spektra hipotetičke sekvence dužine *L* (zbog FFT transformacije taj skup je određen vrednošću prvog stepena dvojke većeg od *L*, što je dužina spektra). Sekvenca se redom skenira od prve kiseline tako što se uzima podsekvenca dužine *L*, izračuna se: njen ISM spektar, amplituda i signal/šum spektra na

zadatoj frekvenci. Početak podsekvence se, redom, pomera od prve pozicije do poslednje. Rezultat čine podsekvence sa svojim početnim pozicijama u originalnoj sekvenci i vrednosti amplituda na frekvenci *F*, koje se sortiraju u opadajućem redosledu prema vrednosti amplituda.



Slika 4.2.12. Prozor programa *SetSkener*.

Algoritam:

Ulaz: sekvenca *S*, dužina podsekvence *L*, frekvenca *F*.

Izlaz: tabela pozicija podsekvenci dužine *L* sa amplitudama ISM spektra na frekvenci *F*.

1. Za svaku poziciju *i* = 1 do *N*-*L*+1 u sekvenci *S*, gde je *N* dužina sekvence *S*:
   a. Pronaći podsekvencu *T* iz sekvence *S* od pozicije *i* dužine *L*.

b. Izračunati ISM spektar *X* podsekvence *T* i amplitudu *A(i)* spektra *X* na frekvenci *F*.

2. Sortirati pozicije podsekvenci prema vrednosti amplituda *A(i)*.

4.2.6. Modul za procenu biološkog efekta mutacija

Modul za procenu biološkog efekta mutacija se sastoji iz programa AKSkener i LPLPrikaz. Funkcija modula je procena koliko svaka pojedinačna mutacija i kombinacija više mutacija, iz skupu mutacija nekog proteina, utiču na određenu biološku funkciju tog proteina.

4.2.6.1. Program *Mutacije*

Svrha:

- Procenjuje uticaj pojedinačnih mutacija i kombinacija više mutacija na biološku funkciju proteina.

Ulaz:

- Proteinska sekvenca, koja se učitava iz datoteke ili se unosi ručno.
- Jedna ili dve karakteristične frekvence (*F1* i *F2*).
- Lista supstitucija.
- Maksimalan broj istovremenih supstitucija (*N*).

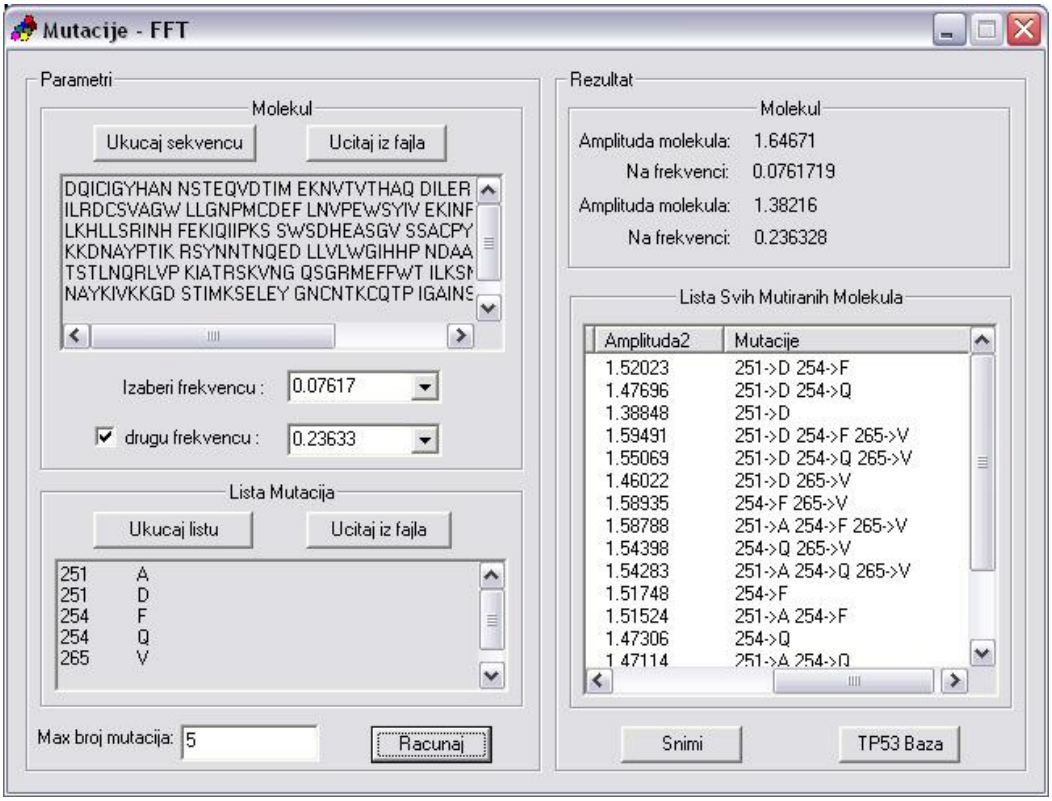
Izlaz:

- Tabela mutacija i vrednosti procenta promene amplitude na zadatim frekvencama, sortirana kao pod c. (dole u opisu). Rezultat se može i snimiti u datoteku gde su sekvence sortirane na sva 4 načina.
- Interval vrednosti amplituda ISM spektra svih mutiranih sekvenci.

Za ulazni protein prvo se generiše lista frekvenci prema rezoluciji spektra koji se, u slučaju FFT, proširuje do stepena dvojke iz koje se selektuju jedna ili dve karakteristične frekvence. Lista supstitucija se unosi u formatu parova pozicija i

aminokiselina, gde aminokiselina može biti i specijalan simbol: „-“ koji se interpretira kao delecija (odsecanje), ili simbol „~“ koji označava stop kodon. U tim slučajevima se signal produžava do dužine originalne sekvence zbog očuvanja rezolucije spektra. Lista se može snimiti u datoteku tipa *.mut i ponovo učitati. Program generiše sve moguće mutirane sekvence iz originalne, ubacivanjem jedne do *N* supstitucija iz liste za koje izračunava ISM spektre primenom FFT transformacije. Sekvence se sortiraju u opadajućem redosledu prema amplitudama jedne (ili dve ako je zadata druga) frekvence ISM spektra na 4 načina:

a. po amplitudama na prvoj frekvenci
b. po amplitudama na drugoj frekvenci
c. po maksimumu amplitude na obe frekvence
d. prema minimumu amplitude na obe frekvence



Slika 4.2.13. Prozor programa *Mutacije*.

Algoritam:

Ulaz: sekvenca *S*, frekvenca *F1*, frekvenca *F2* (opciono), lista supstitucija *Sub*, maksimalan broj supstitucija *M*.

Izlaz: sortirana tabela mutacija prema amplitudama na frekvenci *F1* (i *F2*) u spektrima mutiranih sekvenci.

1. Generisati listu svih mogućih mutacija *Mut* tako da se svaka mutacija sastoji od 1 do *M* različitih supstitucija iz ulazne liste supstitucija *Sub*. Broj mogućih mutacija je $\sum_{i=1}^{M} C_i^K = \sum_{i=1}^{M}\left(\frac{K!}{i!(K-i)!}\right)$, gde je K broj ulaznih supstitucija *Sub*
2. Za svaku mutaciju Mut_i:
   a. Generisati mutiranu sekvencu *Sm* ubacivanjem mutacije Mut_i u sekvencu *S*.
   b. Izračunati ISM spektar *X* sekvence *Sm*, amplitudu *A1*(*i*) spektra *X* na frekvenci *F1* (i amplitudu *A2*(*i*) spektra *X* na frekvenci *F2*).
3. Sortirati listu mutacija *Mut* prema amplitudama *A1*(*i*) (i *A2*(*i*)).

Potprogram *TP53baza* programa *Mutacije*

Svrha:

- Poseban potprogram za analizu TP53 proteina. Nakon analize genskih (standardnih) mutacija u osnovnom delu programa *Mutacije* i detektovanog intervala amplituda na frekvenci 0.2793, analizira se baza *missens* mutacija na proteinu TP53.

Ulaz:

- Tabela nestandardnih mutacija proteina TP53 preuzeta sa URL adrese: http://www.iarc.fr/p53/ [202].

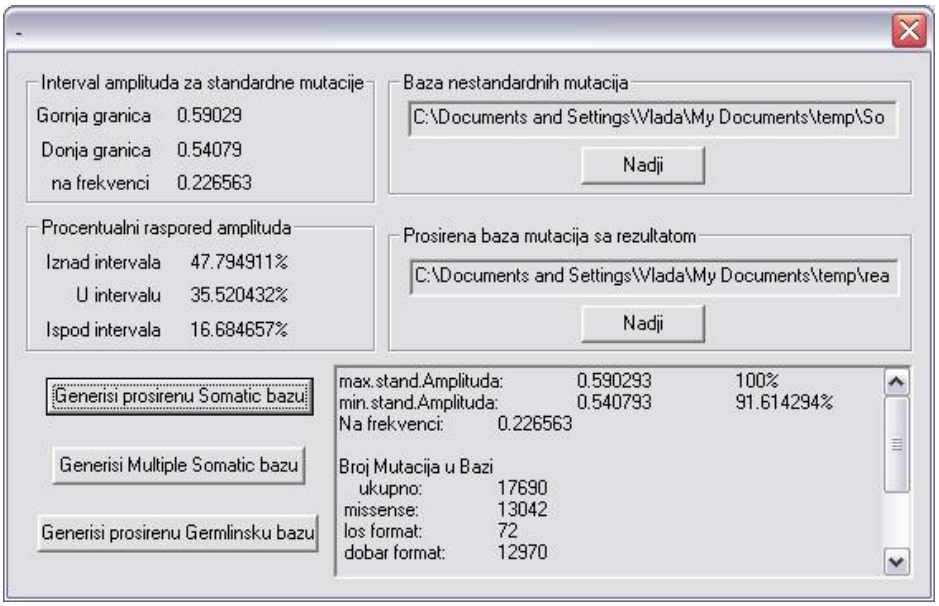
Izlaz:

- Proširena tabela mutacija sa poljima: vrednost amplitude, procenat prema maksimalnoj amplitudi za standardne mutacije, položaj prema intervalu.
- Statistika o broju mutacija, procenat grupisanja mutacija prema intervalu standardnih amplituda.

Iz ulazne baze nestandardnih mutacija redom se primenjuju samo *missens* mutacije (mutacije promene smisla) na osnovni model TP53 proteina i računa se amplituda ISM spektra modifikovane sekvence na frekvenci 0.2793 [203]. Zatim se

sortiraju sekvence u opadajućem redosledu prema procentu promene amplitude i grupišu po položaju u odnosu na interval za standardne mutacije: ispod, između, iznad. Program obrađuje bazu na tri načina:

1. Generisanje proširene baze somatskih mutacija.
2. Generisanje baze višestrukih somatskih mutacija - posmatranje više različitih supstitucija na istom uzorku (*sample_ID*) kao jednu mutaciju.
3. Generisanje proširene baze germinativnih mutacija - grupisanje supstitucija za istu individuu (*Individual_ID*) u jednu mutaciju.



Slika 4.2.14. Prozor dijaloga *TP53* programa *Mutacije*.

4.2.6.2. Program *MutacijeDFT*

MutacijeDFT je varijanta programa *MutacijeFFT*, koja ima istu funkciju i svrhu kao program *MutacijeFFT*, ali umesto FFT transformacije koristi DFT transformaciju, pa se sekvenca ne produžava do stepena dvojke. U specijalnim slučajevima supstitucija: delecije (označava se simbolom „-“) i stop (označava se simbolom „~“), signal se produžava zbog očuvanja rezolucije spektra sa spektra originalne sekvence. Kako je DFT dosta sporija od FFT transformacije, ovaj program se koristi u ograničene svrhe za manje skupove mutacija. Modul *TP53baza* nije implementiran u ovoj varijanti.

4.2.6.3. Program *LPLPrikaz*

Svrha:

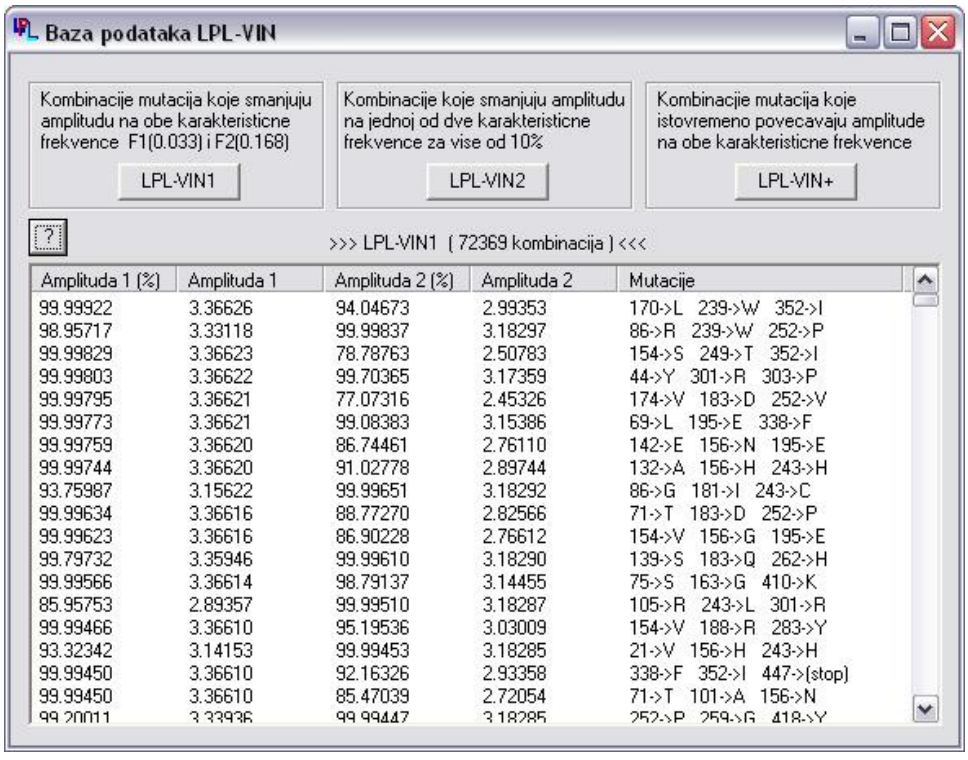
- Učitavanje LPL baze proteina i mutacija, koja je rezultat programa *Mutacije*, i prikazivanje u tabelarnom obliku sa informacijama o vrednosti amplituda i mutacijama, sortiranim u opadajućem poretku prema zbiru amplituda na dve karakteristične frekvence *F1(0.033)* i *F2(0.168)*.

Ulaz:

- Baza LPL proteina sa mutacijama, sastavljena od tri podbaze.

Izlaz:

- Tabela LPL proteina sa mutacijama i amplitudama.



Slika 4.2.15. Prozor programa *LPLPrikaz*.

Baza podataka LPL-VIN sadrži podatke o uticaju pojedinačnih mutacija i njihovih kombinacija na frekventne komponente F(0.031) i F(0.168) u informacionom spektru lipoproteinske lipaze (LPL). Nastala je na osnovu bioinformatičkog kriterijuma

i raspoloživih literaturnih podataka o mutacijama u LPL molekulu pomoću programa *Mutacije*.

LPL-VIN je podeljena u 3 podbaze:

1) LPL-VIN1 koja sadrži kombinacije mutacija koje smanjuju amplitudu na obe karakteristične frekvence.
2) LPL-VIN2 koja sadrži kombinacije mutacija koje smanjuju amplitudu na jednoj od dve karakteristične frekvence za više od 10%.
3) LPL-VIN+ koja sadrži mutacija i njihove kombinacije koje istovremeno povećavaju amplitude na frekvencama F(0.033) i F(0.168).

Baza LPL-VIN predstavlja osnovu prognostičkog testa za kardiovaskularne bolesti zasnovanog na detekciji mutacija na LPL genu i proceni njihovog zbirnog uticaja na funkciju ovog molekula.

4.2.7. Modul za modifikaciju proteina i dizajniranje peptida

Modul za modifikaciju proteina i dizajniranje peptida se sastoji iz programa: Inverz, Kombinator i KombCitac. Modifikacija i dizajniranje proteina se realizuje na osnovu zadatih spektralnih karakteristika proteina.

4.2.7.1. Program *Inverz*

Svrha:

- Generisanje malih sekvenci peptida koji imaju određene biološke karakteristike zadate ISM spektrom.

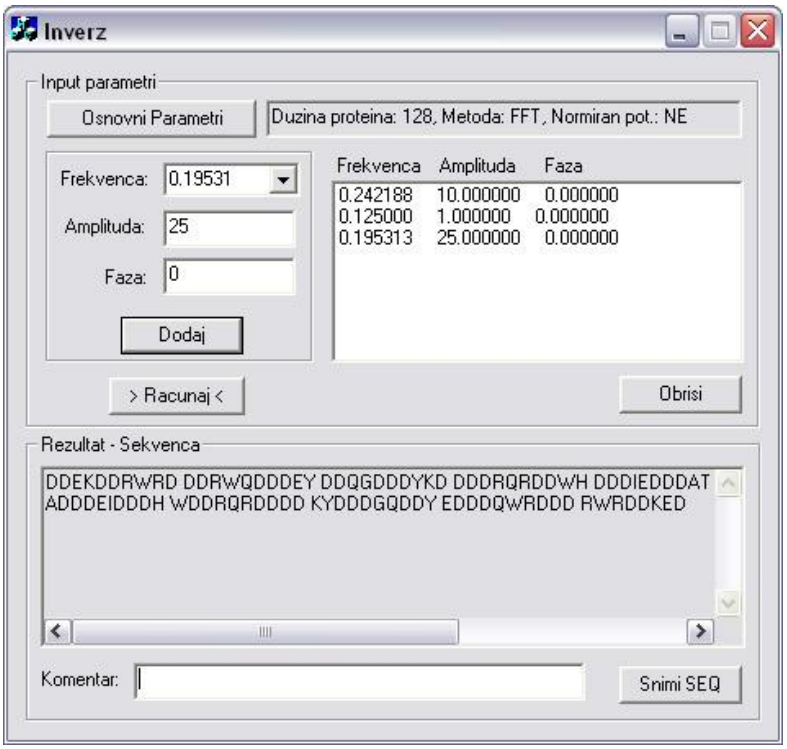
Ulaz:

- Osnovni parametri
    - Dužina proteina
    - Transformacija FFT ili DFT
- Spektar definisan listom pikova koji su zadati parametrima: frekvenca, amplituda i faza.

Izlaz:

- Proteinska sekvenca

Program generiše protein čiji ISM spektar ima vrednosti najbliže zadatom spektru definisanom ulaznim parametrima: lista pikova sa njihovim vrednostima, izabrana metoda transformacije DFT ili FFT i dužinu tražene proteinske sekvence. Program, primenom inverzne Furijeove transformacije na spektru koji karakteriše traženu biološku funkciju proteina, generiše EIIP signal na osnovu koga dekodira sekvencu koja najbliže odgovara tom signalu. Rezultat se sa tekstualnim opisom može snimiti u *.seq datoteku.



Slika 4.2.16. Prozor programa *Inverz*.

Algoritam inverzne transformacije spektra u sekvencu:

Ulaz: vektor *V* zadat listom trodimenzionalnih elemenata (frekvenca, amplituda, faza), dužina sekvence *N*, transformacija DFT ili FFT

Izlaz: proteinska sekvenca *P*.

1. Konstruisati vektor *V* dužine *N* uz pomoć date liste tačaka tako što se ostale koordinate postave na nulu. U slučaju FFT-a proširiti signal na dužinu do najbližeg stepena dvojke. Zatim proširiti realan vektor *V* na kompleksan vektor *W* dodeljivanjem imaginarnom vektoru vrednosti nula.
2. Generisati signal *S* zadatom inverznom transformacijom nad vektorom W. U slučaju FFT, skratiti signal *S* na prvih *N* vrednosti.
3. Dekodirati signal *S* u proteinsku sekvenci *P* preko tabele EIIP vrednosti nalaženjem aminokiselina iz tabele EIIP kodova sa najbližim vrednostima u signalu *S*.

4.2.7.2. Program *Kombinator*

Svrha:

- Konstruisanje proteina sa već poznatim regionima, gde se nepoznati deo generiše tako da se maksimizuju određene karakteristike i funkcije.

Ulaz:

- Dužina proteina
- Dve karakteristične frekvence
- Za svaku poziciju u sekvenci: lista mogućih aminokiselina

Izlaz:

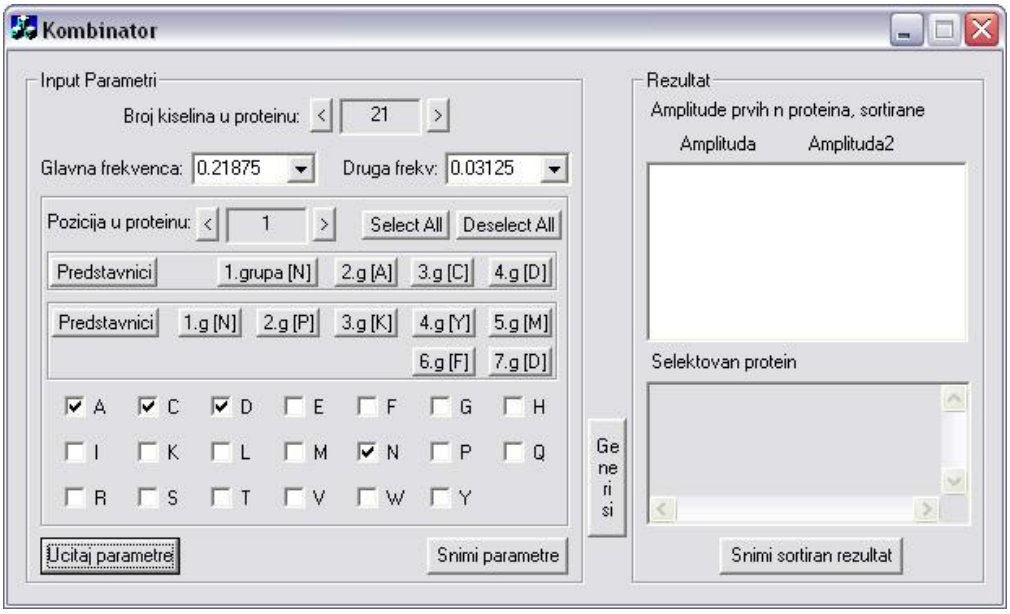
- Tabela proteina sortiranih prema amplitudama na zadatim frekvencama spektra. Rezultat se može snimiti u datoteku tipa *.dat u formatu pogodnim za unos u program *KombCitac*.

Program generiše sve moguće sekvence zadate dužine iz skupa svih mogućih kombinacija definisanih mogućim kiselinama po pozicijama i filtrira one sekvence čija su prva dva pika ISM spektra na traženim karakterističnim frekvencama. Tabela sekvenci se sortira prema amplitudi na prvoj frekvenci, a ako su one iste onda se sortira prema amplitudi na drugoj frekvenci. Zbog ogromnog broja kombinacija (u najgorem slučaju 20^N gde je *N* broj nepoznatih pozicija u peptidu, kada su označene sve kiseline),

izbor mogućih kiselina na poziciji je optimizovan grubljom procenom, uzimanjem predstavnika grupa, gde su grupe kiselina definisane prema sličnim EIIP vrednostima (slika 3.12). Uzete su dve podele na grupe:

1) 4 grupe sa predstavnicima N, A C, D.
2) 7 grupa sa predstavnicima N, P, K, Y, M, F, D.

Posle prve iteracije i grublje procene preko EIIP predstavnika, za finije određivanje aminokiselina, u drugoj iteraciji se na svakoj poziciji selektuju svi članovi detektovane EIIP grupe. Izabrane parametre je moguće snimiti i učitati u datoteku tipa *.kmb.



Slika 4.2.17. Prozor programa *Kombinator*.

Algoritam:

Ulaz: dužina sekvence *N*, frekvenca *F1*, frekvenca *F2*, niz listi aminokiselina *Kis* dužine *N*.

Izlaz: sortirana tabela sekvenci prema amplitudama na frekvencama *F1* i *F2* u informacionim spektrima generisanih sekvenci.

1. Iterativno se generiše sekvenca *S* na osnovu niza listi mogućih kiselina *Kis* po svim pozicijama: Za svaku poziciju *i* od 1 do *N*, bira se kiselina S_i iz liste Kis_i.

Broj mogućih sekvenci je $\prod_{i=1}^{N}|Kis_i|$, gde je $|Kis_i|$ broj mogućih kiselina na poziciji i.

Sek=∅. Za svaku generisanu sekvencu *S*:

a. Izračunati ISM spektar *X* sekvence *S*, amplitudu *A1*(*i*) spektra *X* na frekvenci *F1* i amplitudu *A2*(*i*) spektra *X* na frekvenci *F2*.
b. Ako su prva dva pika spektra *X* na frekvencama *F1* i *F2*, ubaciti u listu *Sek* sekvencu *S* sa amplitudama *A1*(*i*) i *A2*(*i*).

2. Sortirati listu sekvenci *Sek* prema amplitudama *A1*(*i*) i *A2*(*i*):
$\forall\ S_i,S_j \in Sek: S_i > S_j \Leftrightarrow (A1(i)>A1(j))\vee(A1(i)=A1(j) \wedge A2(i)>A2(j))$

4.2.7.3. Program *KombCitac*

Svrha:

- Prikaz rezultata programa *Kombinator*, korišćenjem dva načina sortiranja: prema amplitudama i na osnovu odnosa signal/šum.

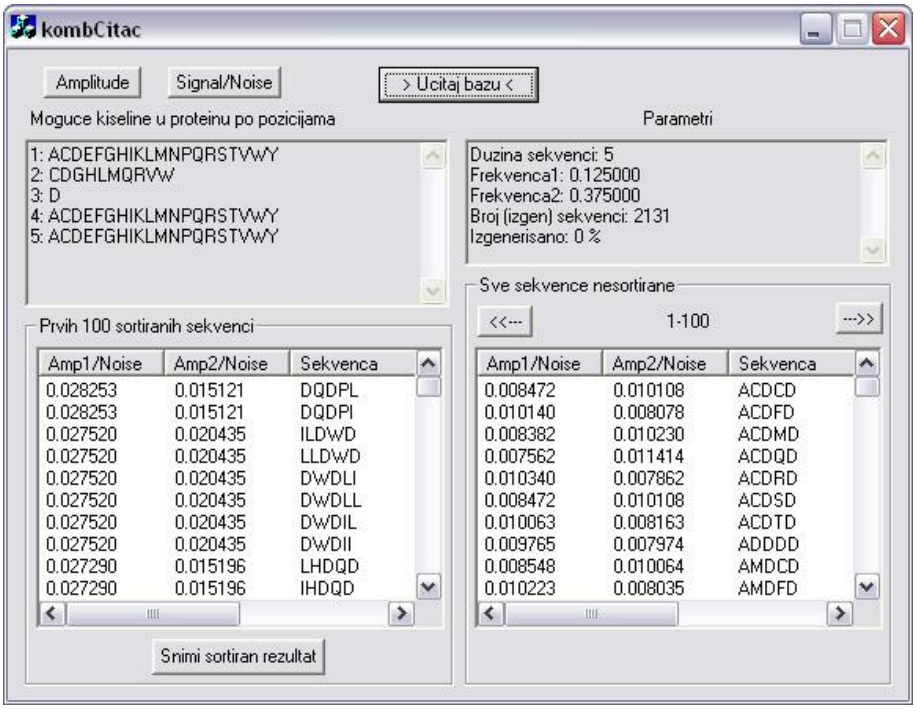
Ulaz:

- Datoteka *.dat koja je izlaz programa *Kombinator*

Izlaz:

- Tabela generisanih sekvenci programom *Kombinator*. Tabelu je moguće snimiti u tekstualnu datoteku.

Program učitava bazu generisanih sekvenci programom *Kombinator* i ispisuje tabelu sekvenci sa vrednostima amplituda i odnosa signal/šum na datim frekvencama. Tabela sekvenci je sortirana u opadajućem redosledu po vrednostima amplituda (signal/šum) na sledeći način: prvo se upoređuju amplitude (signal/šum) na prvoj frekvenci, a ako su jednake vrednosti, onda se upoređuju i na drugoj frekvenci.



Slika 4.2.18. Prozor programa *KombCitac*.

4.2.8. Modul za filogenetsku analizu

Modul za filogenetsku analizu omogućava:

i. Izdvajanje (detekciju) klastera i funkcionalnih grupa u okviru familije proteina ili DNK sekvenci.
ii. Detekciju mutacija bitnih za razdvajanje grupa.
iii. Praćenje evolucije određene biološke funkcije proteina u okviru familije proteina.

Modul sadrži programe ISMStablo i ISMGraf.

4.2.8.1. Program *ISMStablo*

ISMStablo je program za filogenetsku analizu zasnovan na metodi informacionih spektara. Program omogućava detekciju klastera i funkcionalnih grupa u okviru familije proteina, a sa tim i detekciju bitnih mutacija karakterističnih za određene grupe i krucijalnih za razvoj određene biološke funkcije proteina. Program takođe služi

za praćenje filogeneze biološke funkcije proteina u okviru familije proteina. Pored toga program omogućava jednostavan vizuelni prikaz i manipulaciju filogenetskim stablima kroz paletu funkcija programa, kao i markiranje bojama određenih grupa pogodnih za jasnu vizuelnu distinkciju grupa.

Ulaz:

- Baza proteinskih ili DNK sekvenci u datoteci FASTA formata, sa parametrima za računanje rastojanja:
    - *Frekvencna formula* (FF) je definisana regularnim izrazom:
      (['+'|'-'] *K*? (*F* | *F* '/' *Q* | *S*)) +
      K koeficijent, *F*,*Q* frekvence, *S* izabrana metrika na celom spektru:
      S = ['M' |'E' |'C'|'U'|'K'|'P'|'R'|'B'].
    - Frekventni interval.
    - Način selekcije amplitude u frekventnom intervalu: najbliža frekvenca zadatoj, maksimalna u zadatom intervalu oko zadate frekvence, ili srednja vrednost u intervalu.
    - Vrsta rastojanja: razlika po vrednosti amplitudama ili po vrednosti signal/šum.
    - Transformacija: DFT ili FFT.
    - Produžavanje signala: ako je selektovana ova opcija, svi signali se produžuju do zadate dužine ili do dužine najduže sekvence. Ako je zadata metrika za ceo spektar, svi se obavezno produžuju do dužine najduže sekvence iz skupa.

  Izabran algoritam za klasterisanje: NJ ili UPGMA.
- Zapis filogenetskog stabla u datoteci u *Newick* formatu (*.nwk) i *Phylip* formatu (*.tree).
- Zapis topologije stabla bez dužina grana u *Newick* formatu (*.nwk) ili *Phylip* formatu (*.tree).
- Matrica filogenetskih rastojanja između proteina u *Phylip* formatu, sa izabranim algoritmom za klasterisanje

Izlaz:

- Zapis filogenetskog stabla u *Newick* formatu (*.nwk) i *Phylip* formatu (*.tree).
- Zapis topologije filogenetskog stabla bez dužina grana u *Newick* formatu (*.nwk) ili *Phylip* formatu (*.tree).
- Matrica filogenetskih rastojanja između svaka dva proteina u *Phylip* formatu.
- Slika filogenetskog stabla se može snimiti u datoteke tipa *.bmp, *.emf i *.wmf.
- Tekstualna prezentacija stabla se može snimiti u datoteku tipa *.txt.

Pomoću frekvencne formule (FF) definiše se način računanja rastojanja između dve proteinske sekvence. To je linearna kombinacija članova, gde svaki član može biti:

(A) Razlika amplituda informacionih spektara dva proteina na zadatoj frekvenci.

(B) Razlika odnosa dve amplitude na dve zadate frekvence.

(C) Mera na celim spektrima kao vektorima.

Koeficijenti članova su realni brojevi koji daju određenu težinu svakom članu u zbiru. Ako nije dat koeficijent, podrazumevana vrednost je 1.

Metrike na celim spektrima mogu biti [204]:

- Menhetn rastojanje L_1 [205, 206] - **M**

$$d_M(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)$$

- Euklidsko rastojanje L_2 [207] - **E**

$$d_E(X,Y) = \frac{1}{N}\|X - Y\| = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(X_i - Y_i)^2}$$

- Čebiševljevo rastojanje L_∞ [208] - **C**

$$d_C(X,Y) = \max_i(|X_i - Y_i|)$$

- Ugaono rastojanje [209] - **U**

$$d_U(X,Y) = \cos^{-1}\left(\frac{X \cdot Y}{\|X\|\|Y\|}\right) = \cos^{-1}\left(\frac{\sum_{i=1}^{N} X_i Y_i}{\sqrt{\sum_{i=1}^{N} X_i^2}\sqrt{\sum_{i=1}^{N} Y_i^2}}\right)$$

- Kosinusna mera različitosti [210] - **K**

$$d_K(X,Y) = 1 - \cos(X,Y) = 1 - \frac{X \cdot Y}{\|X\|\|Y\|} = 1 - \frac{\sum_{i=1}^{N} X_i Y_i}{\sqrt{\sum_{i=1}^{N} X_i^{\,2}} \sqrt{\sum_{i=1}^{N} Y_i^{\,2}}}$$

- Pirsonovo rastojanje [211] - **P**

$$d_P(X,Y) = 1 - \frac{\sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}}, \quad \overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

- Kanbera rastojanje (eng. *Canberra*) [212] - **R**

$$d_R(X,Y) = \frac{1}{N} \sum_{i=1}^{N} \frac{|X_i - Y_i|}{|X_i| + |Y_i|}$$

- Braj-Kurtis mera različitosti (eng. *Bray-Curtis*) [213] - **B**

$$d_B(X,Y) = \frac{1}{N} \frac{\sum_{i=1}^{N} |X_i - Y_i|}{\sum_{i=1}^{N} |X_i + Y_i|}$$

Neke od metrika su normirane sa vrednošću 1/*N* zbog usaglašavanja mere na različitim skupovima sekvenci sa različitim brojem sekvenci, gde je *N* broj sekvenci u skupu. Pošto je za analizu određenog skupa sekvenci *N* konstantna vrednost, sledi da je 1/*N* konstanta skaliranja mere koja nema uticaja na validnost metrika.

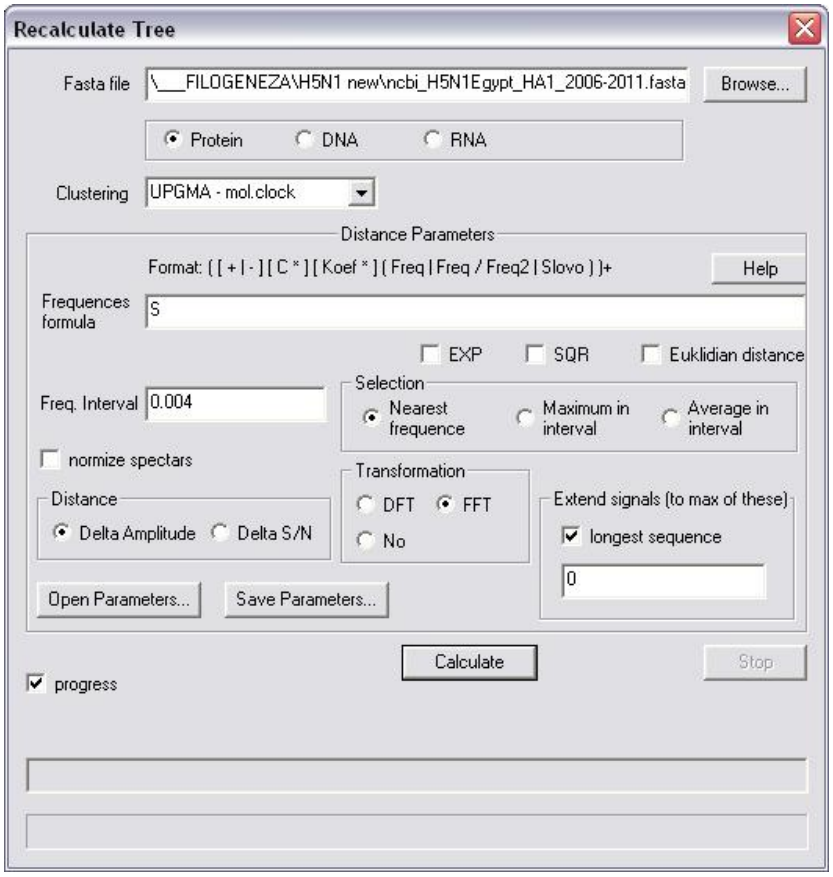
Korišćenje i testiranje programa pokazuje da se po sličnosti u rezultujućim stablima grupišu metrike: (a) K,P ; (b) M,E,C,U,B ; (c) R daje loše rezultate.

Na osnovu ulazne frekvence formule, rastojanje između sekvenci *X* i *Y* je linearna kobinacija:

$$d(X,Y) = \sum_{i=1}^{L1} (K_i |A_X(F_i) - A_Y(F_i)|) + \sum_{i=1}^{L2} \left(K_i \left| \frac{A_X(F_i)}{A_X(Q_i)} - \frac{A_Y(F_i)}{A_Y(Q_i)} \right| \right) + \sum_{i=1}^{L3} (K_i d_i(S_X, S_Y))$$

gde su K_i koeficijenti članova FF, *L1* broj članova tipa (A), *L2* broj članova tipa (B), *L3* broj članova tipa (C), F_i i Q_i frekvence i-tog člana FF, d_i metrike na vektorima, S_X i S_Y informacioni spektri sekvenci *X* i *Y* redom, $A_X(F)$ amplituda spektra sekvence *X* na frekvenci *F*, i slično $A_Y(F)$.

Prema zadatim ulaznim parametrima, koji definišu kako se računaju spektri sekvenci i rastojanje između sekvenci (spektara), generiše se matrica rastojanja između svake dve sekvence. Zatim se iz matrice generiše filogenetsko stablo primenom metoda klasterisanja UPGMA ili NJ. U slučaju korišćenja metrike na celim spektrima (C) rastojanja ne zadovoljavaju uslov aditivnosti pa je bolje je koristiti UPGMA metoda, dok u slučaju metrike na pojedinačnoj frekvenci (A) i odnosu amplituda (B), rastojanja zadovoljavaju uslov aditivnosti pa NJ metoda daje korektne rezultate.



Slika 4.2.19. Prozor dijaloga za unos ulaznih parametara u program *ISMStablo*.

Prikazivanje stabla u programu *ISMStablo*

Filogenetsko stablo se može prikazati u grafičkom (slike 4.2.22, 4.2.23 i 4.2.24) ili tekstualnom obliku (slika 4.2.25). Opciono se može isključiti/uključiti ispis naziva sekvence, kao i vrednost vezana za sekvencu definisana frekvencnom formulom u slučaju (A) i (B). U naslovu slike se ispisuju vrednosti ulaznih parametara. Vertikalno

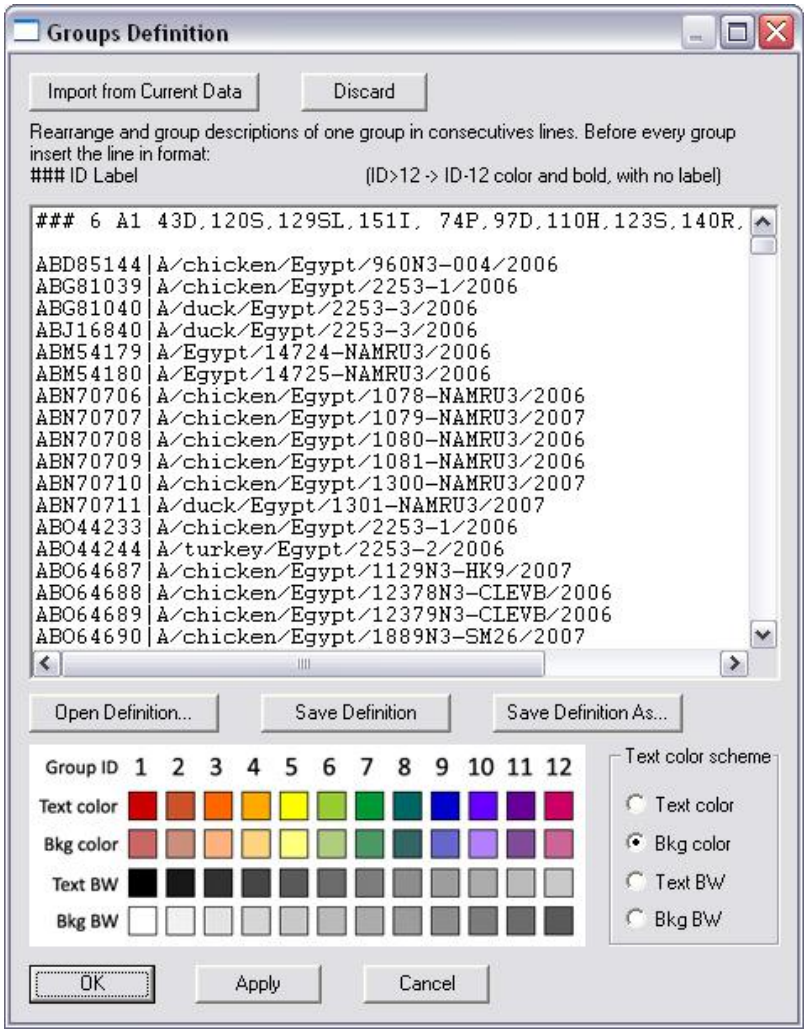
skaliranje stabla omogućava potpun vertikalni prikaz stabla: (i) u punoj razmeri, (ii) procentualno skaliran u odnosu na potpun prikaz, (iii) prilagođen tako da celo stablo stane u prozor programa. Postoje dve vrste oblika stabla: linearni (slika 4.2.22) i kružni (slika 4.2.23). Osim prikaza standardnog stabla sa određenim dužinama grana, postoji i prezentacija topologije stabla nezavisna od dužina grana.

Interaktivna promena uređenja stabla u programu *ISMStablo*

Levi klik mišem na čvor stabla, u zavisnosti od uključene opcije, omogućava: (i) selektovanje novog korena stabla, (ii) simetričnu zamena podgrana, (iii) simetričnu zamena podgrana sa rekurzivnom izmenom u svim podčvorovima. Postoje komande za automatsko sortiranje podgrana čvora tako da stablo bude (i) simetrično, (ii) rastuće u smislu dužina podgrana i to: (ii.a) samo za trenutan čvor ili (ii.b) rekurzivno za sve podčvorove. Trenutno uređenje stabla se može snimiti, a u trenutni prikaz se mogu povratiti originalno ili snimljeno uređenje.

Oznaka bojom grupa sekvenci u programu *ISMStablo*

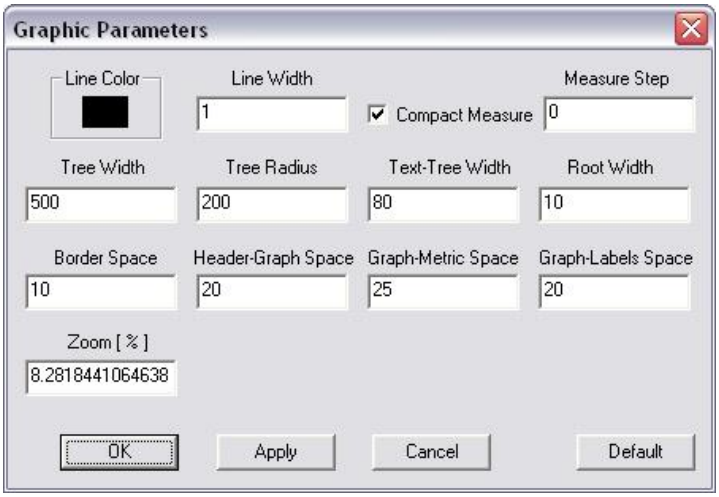
Za jasnije vizuelno izdvajanje grupa sekvenci, omogućeno je bojenje različitim bojama više grupa sekvenci u stablu. Tekstualni format koji definiše bojenje se može sačuvati i učitati iz datoteke sa ekstenzijom *.gdef. Svaka definicija grupe sekvenci počinje novim redom oblika „*$$$ N naziv*“, gde je *N* indeks boje, *naziv* je naziv grupe, ispod koga se dalje ređaju, u svakom redu po jedan, nazivi sekvenci koji pripadaju toj grupi. Postoji dvanaest definisanih različitih boja i četiri vrste bojanja: bojenje teksta naziva sekvenci ili pozadine teksta, i to u boji ili monohromatski.



Slika 4.2.20. Prozor dijaloga za definiciju grupa sekvenci za oznaku bojama u programu *ISMStablo*.

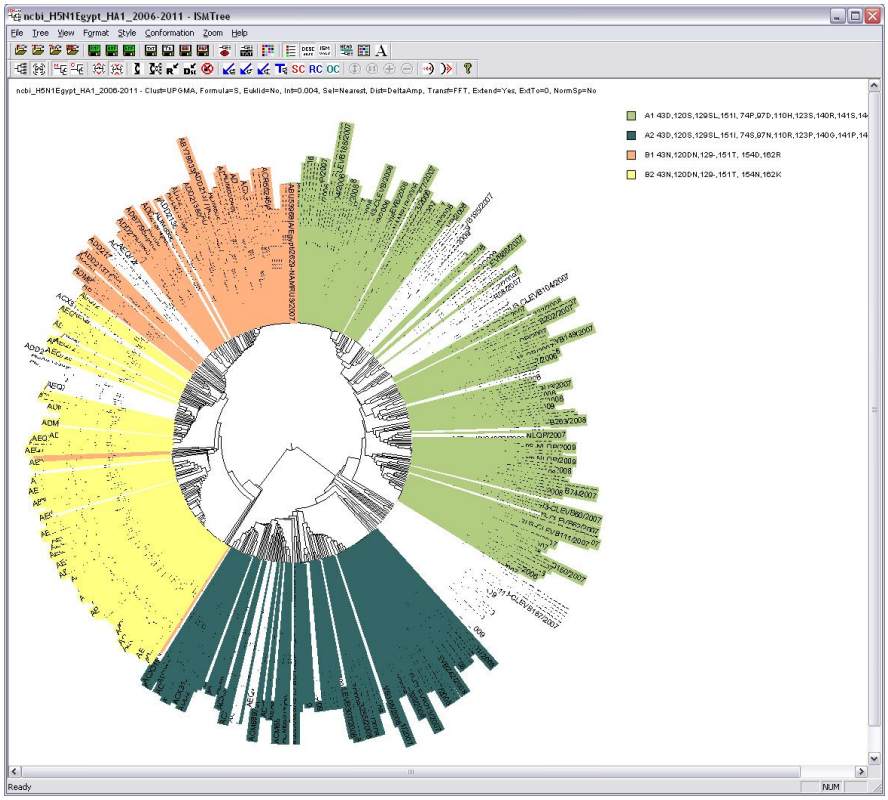
Parametri prikaza stabla u programu *ISMStablo*

Postoji nekoliko parametara koji određuju kako će se iscrtati stablo: boja i debljina linija, tipografsko pismo (eng. *Font*), boja teksta, širina stabla, veličina ivica slike i razmaka između delova slike u pikselima, procenat uveličanja prikaza stabla (eng. *zoom*) po vertikalnoj osi, koji može biti postavljen na dinamičku vrednost tako da celo stablo stane tačno u prozor aplikacije. Oznaka razmere ispod stabla, koja prikazuje meru dužina grana stabla, može se iscrtati u kompaktnom ili izbaždarenom obliku.

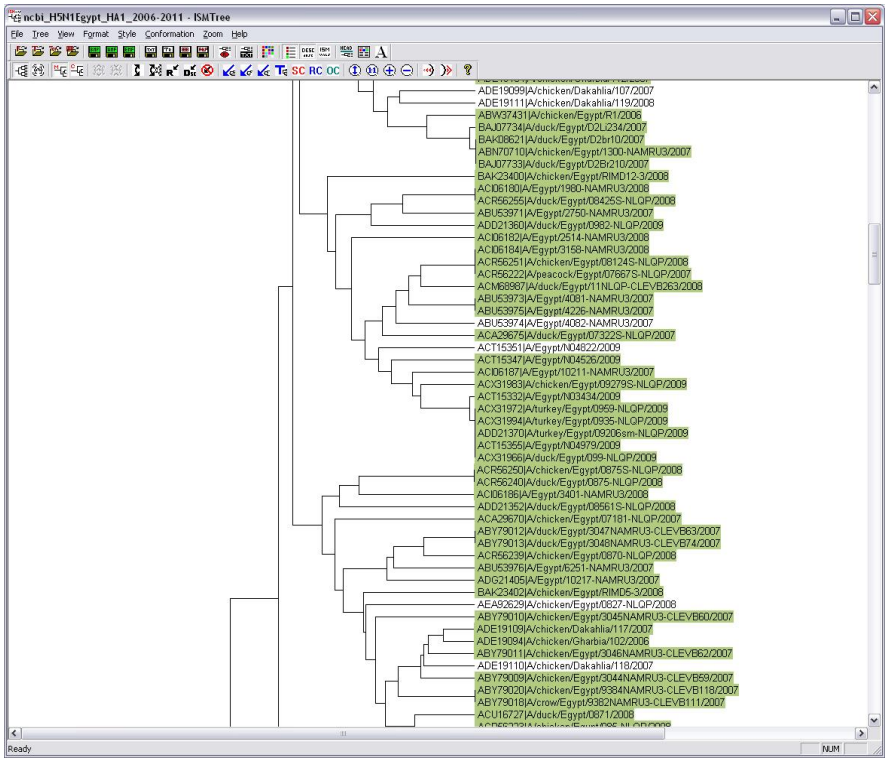


Slika 4.2.21. Prozor dijaloga za unos parametara prikaza u programa *ISMStablo*.

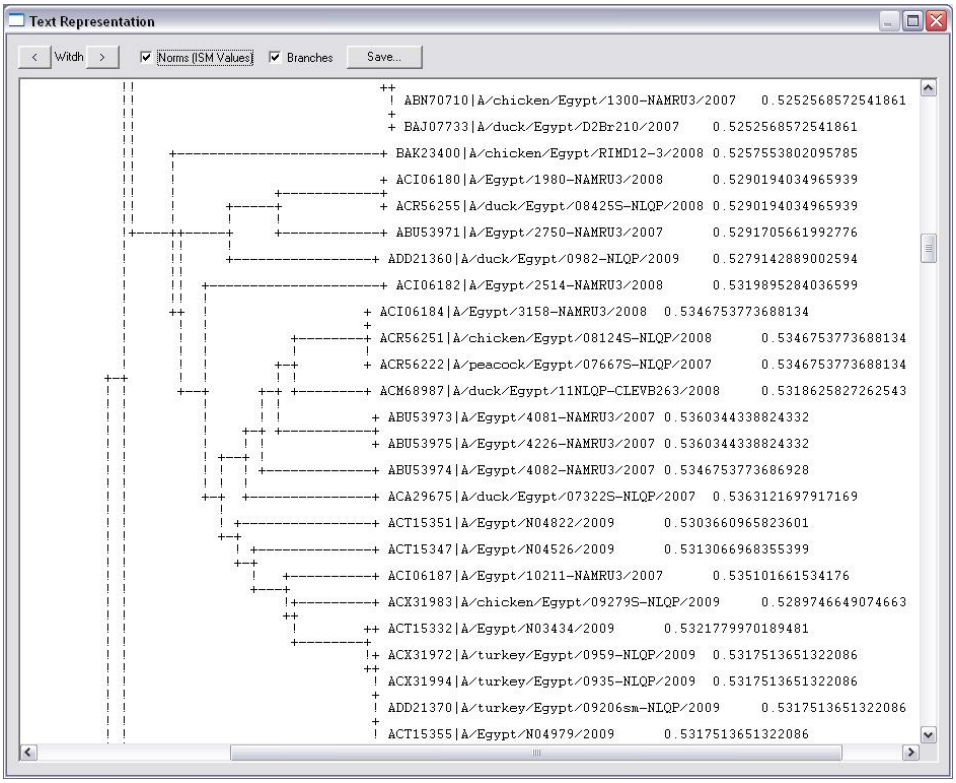
Slika 4.2.22. Prozor programa *ISMStablo*, sa primerom linearnog grafičkog prikaza filogenetskog stabla, vertikalno skalirano tako da celo stablo stane u prozor.



Slika 4.2.23. Prozor programa *ISMStablo*, sa primerom kružnog grafičkog prikaza filogenetskog stabla.



Slika 4.2.24. Prozor programa *ISMStablo*, sa primerom linearnog grafičkog prikaza filogenetskog stabla, sa punim vertikalnim uveličanjem.



Slika 4.2.25. Prozor programa *ISMStablo*, sa primerom tekstualnog prikaza filogenetskog stabla.

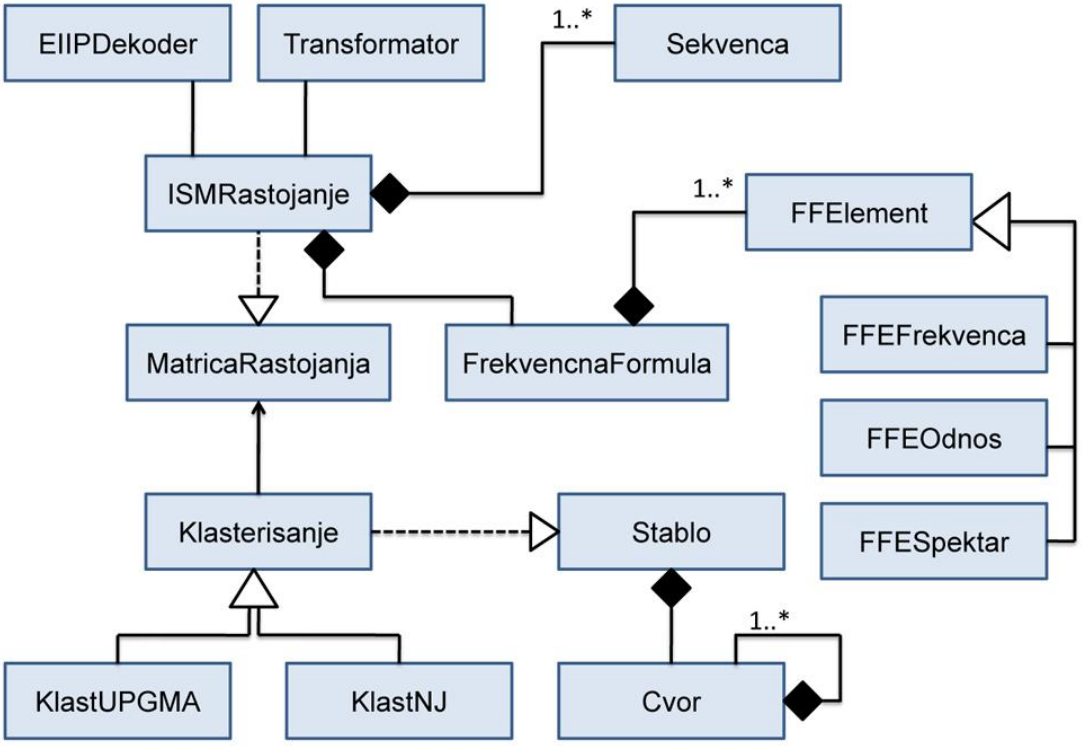
Implementacija programa *ISMStablo*

Osnovne klase implementirane u programu ISMStablo modula za filogenetsku analizu, pored klasa jezgra EIIP/ISM platforme, su:

- *FrekvencnaFormula* ima sledeće funkcije: sintaksno analizira zapis sekvence formule koja definiše ISM rastojanje i kreira elemente formule (*FFElement*) koji mogu biti: (i) rastojanje na pojedinačnoj frekvenci (*FFEFrekvenca*), (ii) odnos amplituda na dve frekvence (*FFEOdnos*) i (iii) mera na celim spektrima (*FFESpektar*);
- *Klasterisanje* kreira objekat *Stablo* od objekta *MatricaRastojanja*, koristeći UPGMA ili NJ algoritam;
- *Stablo* predstavlja filogenetsko stablo i sadrži korenski čvor tipa *Cvor* koji rekurzivno sadrži podčvorove tipa *Cvor*.
- *MatricaRastojanja* sadrži matricu ISM rastojanja između svake dve sekvence;

- *ISMRastojanje* generiše objekat *MatricaRastojanja* na osnovu: (i) objekta *FrekvencnaFormula* u kojem se definiše formula za rastojanje, (ii) skupa sekvenci koje analizira i pamti ih u nizu objekata tipa *Sekvenca* i (iii) objekata jezgra platforme (*EIIPDekoder*, *Transformator*).

Na slici 4.2.26 je prikazan odnos osnovnih klasa u programu ISMStablo.



Slika 4.2.26. Klasni dijagram programa ISMStablo.

4.2.8.2. Program *ISMGraf*

ISMGraf je alat za brzo vizuelno grupisanje skupa proteinskih ili DNK sekvenci zasnovan na ISM metodi. Program omogućava detekciju funkcionalnih grupa u okviru familije proteina, a sa tim i detekciju bitnih mutacija karakterističnih za određene grupe.

Ulaz:

- Baza proteinskih ili DNK sekvenci u datoteci FASTA formata sa parametrima za računanje rastojanja kao u programu *ISMStablo*.
    - Izbor algoritma za konstrukciju grafa, koji može biti:
        - Zasnovan na brzini, sa parametrima: vremenski period za jedan korak *timestep*, masa i *dumping* $\in$ (0,1).
        - Klasičan
        - Zasnovan na metodi simuliranog kaljenja (eng. *Simulated Anealing, SA*), koji je povoljan izbegavanje lokalnog optimuma, sa parametrima: početna temperatura *startT* i koeficijent hlađenja *koefC* $\in$(0,1).

          U metode SA spadaju sledeći algoritmi:
            - *Fruchterman-Reingold* [214]
            - *Walshaw* [215].
            - Efikasan [216]
    - Kriterijum za zaustavljanje algoritma:
        - Minimalan pomeraj (*min_move*)
        - Maksimalan broj iteracija (*max_iterations*)
    - Parametri sila u modelu elektro-opruga:
        - *K* - koeficijent elastičnosti opruge.
        - R_k - elektrostatički koeficijent odbijanja.
- Zapis grafa u datoteci u posebno definisanom formatu (*.graph).
- Matrica filogenetskih rastojanja između proteina u *Phylip* formatu.

Izlaz:

- Zapis grafa u datoteci u posebno definisanom formatu (*.graph).
- Matrica filogenetskih rastojanja između svaka dva proteina u *Phylip* formatu.
- Slika grafa se može snimiti u datoteke tipa *.bmp, *.emf i *.wmf.

ISMGraf je alat za brzu vizuelnu analizu strukture grafa kojim je predstavljen skup proteina sa ISM rastojanjima između čvorova. Kao kod programa *ISMStablo*, ulazni parametri definišu ISM rastojanje između dve sekvence na osnovu kojih se

generiše matrica rastojanja. Zatim se, na osnovu modela *elektro-opruga* (eng. *spring-electro*) za predstavljanje grafa [216-221], gde su dužine grana grafa između svaka dva čvora vrednosti iz matrice rastojanja sekvenci i čvorovi grafa same sekvence, primenjuje izabrani *usmeren silom* (eng. *force-directed*) algoritam za crtanje grafa koji sekvence predstavlja u dvodimenzionalnom vektorskom prostoru (prozoru programa).

Kod modela elektro-opruga, sila između čvorova se sastoji od privlačne sile opruge (Hukov zakon) $F = k\Delta x$, i elektrostatičke sile odbijanja (Kulonov zakon) $F = \frac{1}{4\pi\varepsilon}\frac{q_1 q_2}{r^2}$. Ukupna sila privlačenja između dva čvora se može predstaviti formulom:

$$F_a = K(d_{graf} - d_{mat}) - \frac{R_k}{d_{graf}^{\,2}}$$

gde je K koeficijent opruge, R_k koeficijent odbijanja, d_{graf} realno dinamičko rastojanje između čvorova u dvodimenzionalnom prostoru, d_{mat} rastojanje između sekvenci u N-dimenzionalnom vektorskom prostora modela grafa (zadata matrica rastojanja), gde je N rezolucija informacionih spektara sekvenci.

Algoritam usmeren silom za crtanje grafa pronalazi optimalni raspored čvorova, tako što iterativno računa novi vektor pozicije svakog čvora na osnovu prethodne pozicije, minimizujući energiju sistema modela, dok god nije ispunjen neki od kriterijuma zaustavljanja. U zavisnosti od algoritma, u svakom koraku vektor nove pozicije P_{i+1} svakog čvora se računa po formuli:

- $\overrightarrow{P_{i+1}} = \overrightarrow{P_i} + timstep\,\overrightarrow{V_i}, \quad \overrightarrow{V_{i+1}} = (\overrightarrow{V_i} + timestep\,\overrightarrow{F_a})\,dumping$

  gde je V_i vektor brzine u i-tom koraku, F_a ukupna sila između trenutnog čvora i svih ostalih (algoritam zasnovan na brzini).

- $\overrightarrow{P_{i+1}} = \overrightarrow{P_i} + \overrightarrow{F_a}$

  gde je F_a ukupna sila između trenutnog čvora i svih ostalih u i-tom koraku (klasičan algoritam).

- $\overrightarrow{P_{i+1}} = \overrightarrow{P_i} + \frac{\overrightarrow{F_a}}{\left|\overrightarrow{F_a}\right|}\min(\left|\overrightarrow{F_a}\right|, t_i), \quad t_{i+1} = t_i\, koefC$

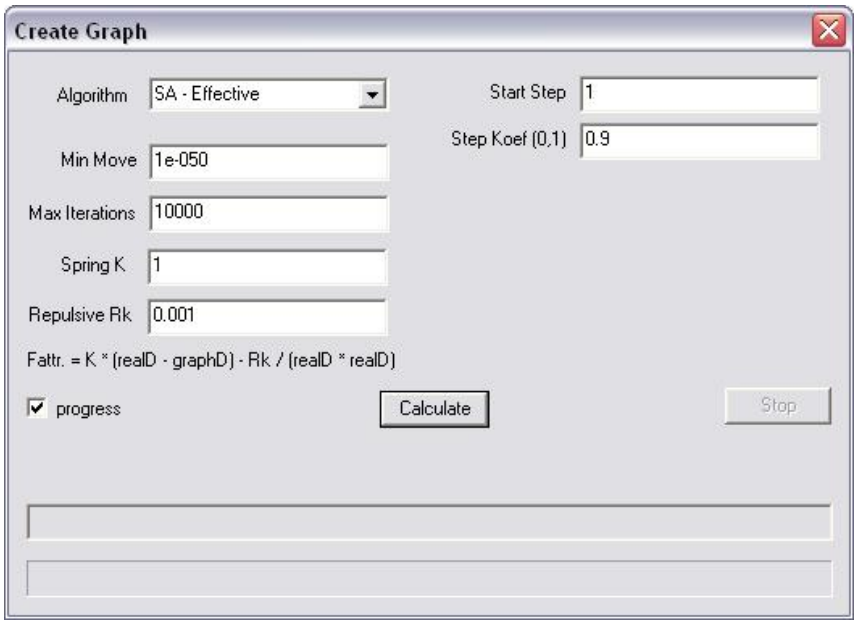
gde je t_i temperatura u i-tom koraku, sa početnom vrednošću *startT* na početku algoritma (*Fruchterman-Reingold* algoritam [214]).

- $\overrightarrow{P_{i+1}} = \overrightarrow{P_i} + t_i \overrightarrow{F_a}, \quad t_{i+1} = t_i \, koefC$

  (*Walshaw* algoritam [215]).

- $\overrightarrow{P_{i+1}} = \overrightarrow{P_i} + t_i \overrightarrow{F_a}, \quad t_{i+1} = \begin{Bmatrix} t_i \, koefC, 5 \mid p_i \\ t_i / koefC, \neg(5 \mid p) \end{Bmatrix}, \quad p_{i+1} = p_i + 1$

  gde je p_i identifikator progresa sa početnom vrednošću nula (Efikasan algoritam [216]).



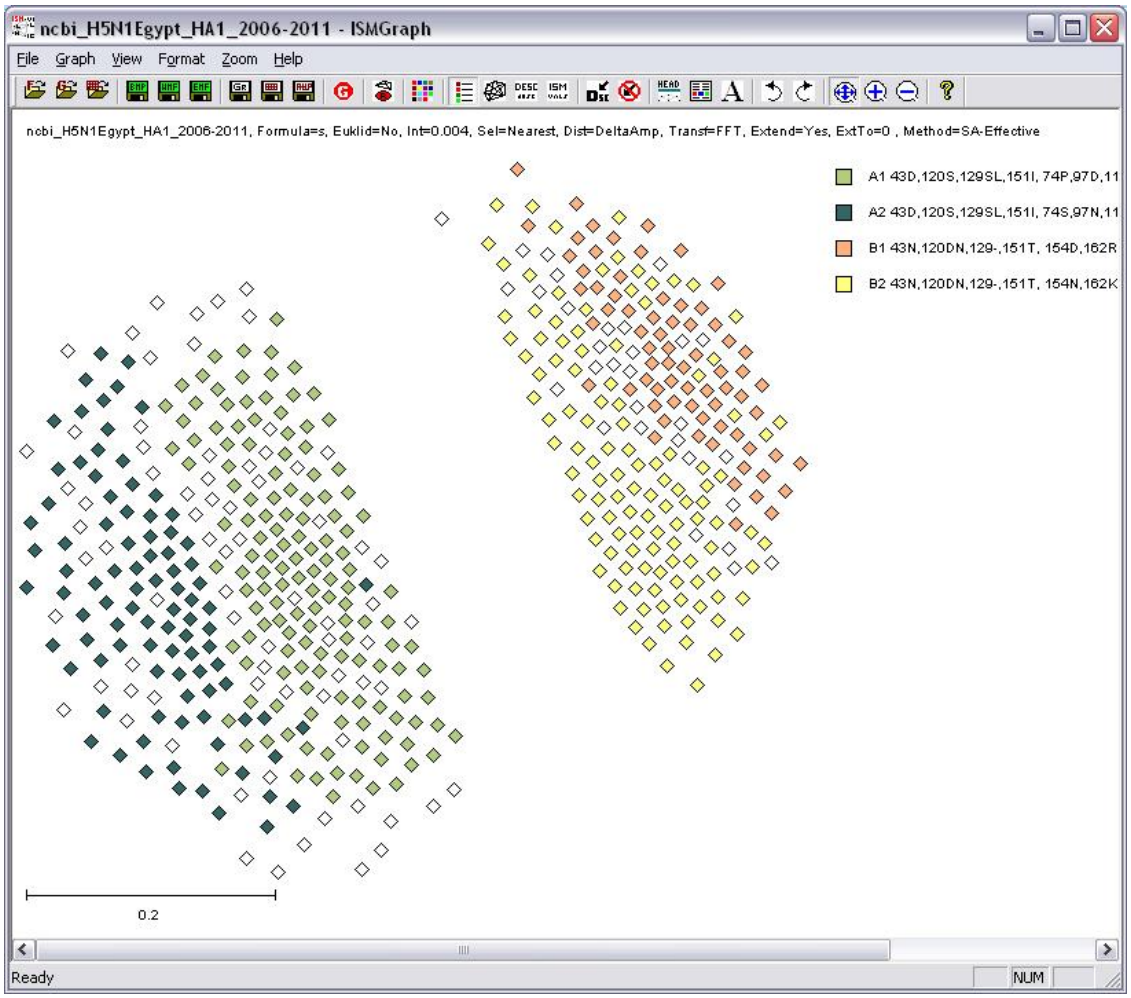
Slika 4.2.27. Prozor dijaloga za izbor algoritma za konstrukciju grafa i unos ulaznih parametara, u programu *ISMGraf*.

Algoritam se zaustavlja kada je ispunjen jedan od uslova: (i) ukupna energija sistema, odnosno ukupan pomeraj u iteraciji je manji od *min_move* ili (ii) broj koraka je veći od *max_iterations*, gde su *min_move* i *max_iterations* zadati parametri.

Prikazivanje grafa u programu *ISMGraf*

Pri iscrtavanju grafa sekvenci u naslovu slike se ispisuju vrednosti ulaznih parametara. Opciono se može isključiti/uključiti ispis naziva sekvence, kao i grana

grafa. Skaliranje grafa omogućava uvećanje u umanjenje prikaza grafa, koje se može prilagoditi tako da ceo graf stane u prozor programa. Posebne komande omogućavaju rotiranje grafa u smeru kazaljke sata i obrnuto za ugao $\pi/16$. Bojenje grupa sekvenci je definisano istim formatom kao u programu *ISMStablo*, a definicija bojenja se može učitati i snimiti u datoteku sa ekstenzijom *.gdef.

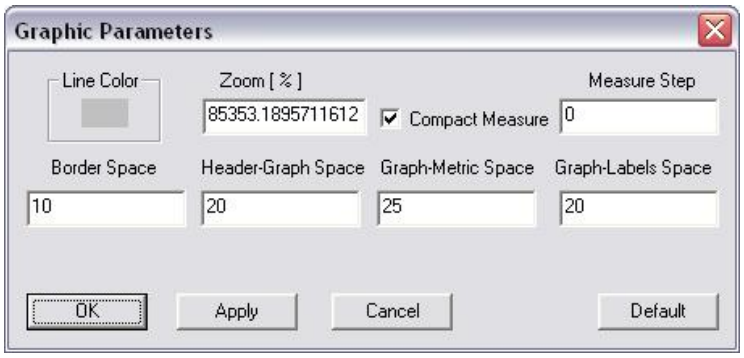


Slika 4.2.28. Prozor programa *ISMGraf*, sa primerom prikaza grafa uz isključenu opciju ispisa naziva sekvenci.

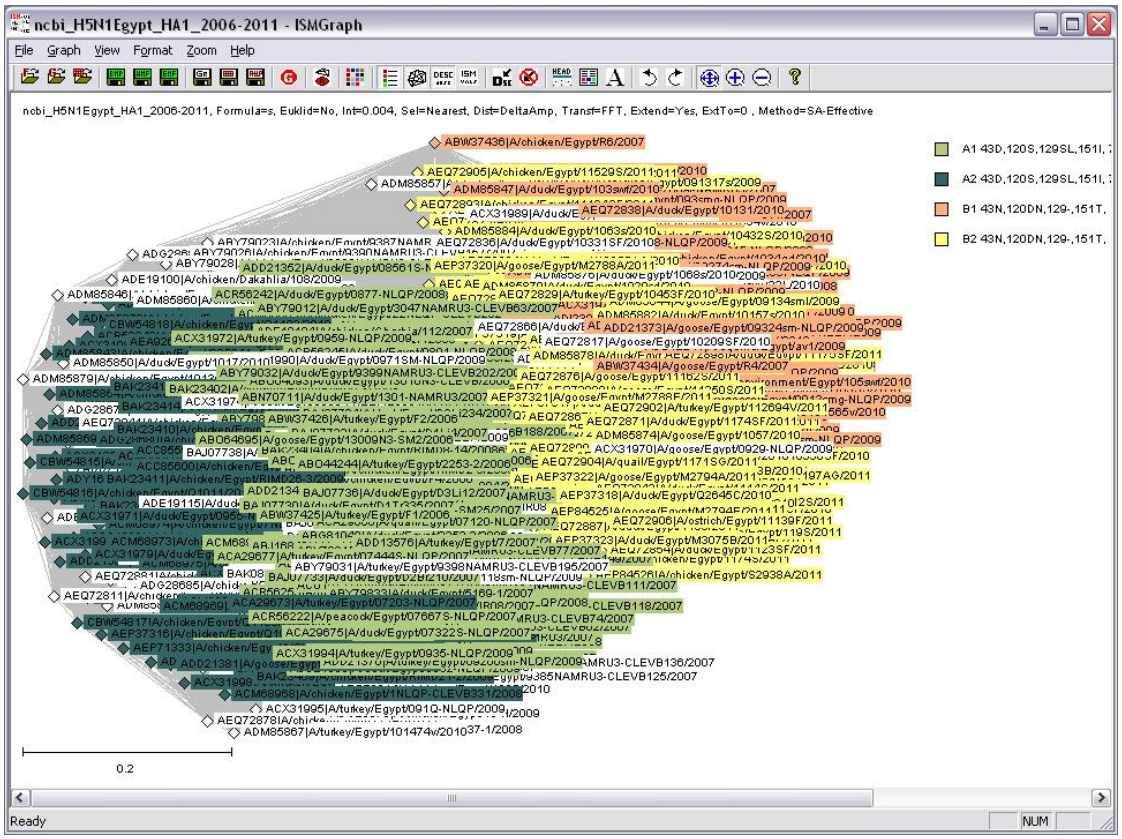
Parametri prikaza grafa u programu *ISMGraf*

Nekoliko parametara, kao i kod programa *ISMStablo*, određuju kako će se iscrtati graf: boja i debljina linija, tipografsko pismo (eng. *Font*) i boja teksta, veličina ivica slike i razmaka između delova slike u pikselima, procenat uveličanja prikaza grafa

(eng. *zoom*), koji može biti postavljen na dinamičku vrednost tako da ceo graf stane u prozor aplikacije. Oznaka razmere ispod grafa, koja prikazuje meru dužina grana grafa, može se iscrtati u kompaktnom ili izbaždarenom obliku.



Slika 4.2.29. Prozor dijaloga unosa parametara za prikaz grafa u programu *ISMGraf*.



Slika 4.2.30. Prozor programa *ISMGraf*, sa primerom prikaza grafa uz uključenu opciju ispisa naziva sekvenci.

4.2.9. Modul za pretraživanje molekulskih biblioteka

Modul za pretraživanje molekulskih biblioteka se sastoji od programa: Chemdb2Alati, NiaidPubchemSpoj, FormulaKalkulator, ValencPotencKalk, PubchemParser, PubchemTxtParser, QSARParser, Raspodela i Raspodela2D. Funkcija modula je pretraživanje molekulskih biblioteka ChemDB i PubChem, računanje EIIP vrednosti molekula i prikazivanje osnovne statistike raspodele EIIP vrednosti jedinjenja iz molekulskih baza.

4.2.9.1. Program *ChemdbAlati*

Svrha:

- Pretraživanje molekulske biblioteke ChemDB i izdvajanje potrebnih informacija o jedinjenjima.
- Sortiranje i filtriranje jedinjenja po klasama kojima pripadaju.

Ulaz:

- Datoteke preuzete sa URL adrese ChemDB servisa za pretragu jedinjenja: http://chemdb.niaid.nih.gov/struct_search/.

Izlaz:

- Datoteka u formatu tabele sa poljima koja su specifična za potprograme.

Program parsira (sintaksno analizira) preuzete *.asp i *.html datoteke i ispisuje informacije u datoteku u tabelarnom formatu (polja su razdvojena *tab* simbolom), sa kolonama koje zavise od potprograma.

Program *ChemdbAlati* se sastoji od nekoliko potprograma koji se razlikuju prema formatu datoteka koje se parsiraju, gde format zavisi od internet izvora, tj. URL adrese sa koje su preuzete i potrebnih informacija koje se izvlače iz tih datoteka.

Potprogram *Parsiraj sve chemdb asp datoteke*

URL adresa preuzetih datoteka koje se parsiraju je:
http://chemdb.niaid.nih.gov/struct_search/all/url_search.asp?aids_no=(*)
Informacije koje se parsiraju, odnosno kolone u izlaznoj tabeli su:
FileName, AIDSNum, ChemicalName, Company, Classes, Anti-HIV Cellular data, Anti-HIV Enzyme data, MW, Formula.

Potprogram *Parsiraj Anti-HIV Cell asp [Cellular-based Details]*

URL adresa preuzetih datoteka koje se parsiraju je:
http://chemdb.niaid.nih.gov/struct_search/ivt/ivt_details.asp?AIDS=AIDSNum
Informacije koje se parsiraju, odnosno kolone u izlaznoj tabeli su:
FileName, AIDS#, EC50, IC50, TI.
Ako ima više vrednosti u jednom ulaznom polju (npr. posle identifikatora *AIDS#*), vrednosti u izlaznoj koloni se odvajaju zadatim separatorom (podrazumevani je " $ ").

Potprogram *Parsiraj Anti-HIV Enzyme asp [Enzyme Inhibition Details]*

URL adresa
preuzetih datoteka koje se parsiraju je:
http://chemdb.niaid.nih.gov/struct_search/ei/ei_details.asp?AIDS=AIDSNum
Informacije koje se parsiraju, odnosno kolone u izlaznoj tabeli su:
FileName, AIDS#, IC50.
Vrednosti su odvojene kao kod potprograma *Parsiraj Anti-HIV Enzyme asp*.

Potprogram *Parsiraj Anti-HIV OI htm [Anti-Opportunistic infection]*

URL adresa preuzetih datoteka koje se parsiraju je:
http://chemdb.niaid.nih.gov/struct_search/oi/oi_details.asp?AIDS=AIDSNum
Informacije koje se parsiraju, odnosno kolone u izlaznoj tabeli su:
FileName, AIDS#, Cytotox IC50, IVT IC50 or EC50, IVT SI.

Potprogram *Prebroj klase*

Kao ulaz, program uzima datoteku koja je rezultat potprograma *Parsiraj sve chemdb2 asp datoteke* i prebrojava koliko jedinjenja ima po klasi, gde jedno jedinjenje može pripadati većem broju klasa. Klase su razdvojene simbolom ";". Potprogram sortira rezultat po klasama i zapisuje ga *Excel* datoteku sa kolonama: *Klasa, Broj jedinjenja*.
Rezultat se može uporediti sa rezultatom sa URL adrese:
http://chemdb2.niaid.nih.gov/struct_search/class/class_search.htm.

Potprogram *Filtriraj po klasi*

Za zadatu klasu i ulaznu datoteku, koja je rezultat potprograma *Parsiraj sve chemdb2 asp datoteke*, potprogram *Filtriraj po klasi* filtrira samo one redove koji sadrže zadatu klasu i ispisuje rezultat u istom formatu i kolonama kao u potprogramu *Parsiraj sve chemdb2 asp datoteke*: *FileName, AIDSNum, ChemicalName, Company, Classes, Anti-HIV Cellular data, Anti-HIV Enzyme data, MW, Formula*.

Potprogram *Parsiraj sve pubchem htm datoteke*

Funkcije potprograma *Parsiraj sve pubchem htm datoteke* su skidanje i filtriranje podataka, odnosno određenih hemijskih jedinjenja sa PubChem internet stranice za kasniju analizu i računanje AQVN i EIIP vrednosti tih jedinjenja.

URL adresa preuzetih datoteka koje se parsiraju je:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=pcsubstance&term=AIDSNum[sourceid],niaid[sourcename]
Informacije koje se parsiraju, odnosno kolone u izlaznoj tabeli su:
FileName, SID, Compound_ID, Source_NIAID (*AIDS#* tj. *AIDSNum*).

Slika 4.2.31. Prozor programa *ChemdbAlati*.

4.2.9.2. Program *NiaidPubchemSpoj*

Svrha:

- Univerzalno spajanje i uparivanje rezultata pretraživanja NIAID i PubChem baza po zadatim kolonama i dodatnim uslovima.

Ulaz:

- NIAID baza koja je rezultat programa *Chemdb2Alati* (*Parsiraj sve chemdb2 asp fajlove*) i redni broj *AIDSNum* kolone.
- PubChem baza sa rednim brojem *NIAID_ID* kolone, gde je baza dobijena na sledeći način:
    1. Za svaku *AIDSNum* vrednost iz NIAID baze preuzete su *.htm datoteke vezane za tu vrednost: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=pcsubstance&term=AIDSNum[sourceid],niaid[sourcename]

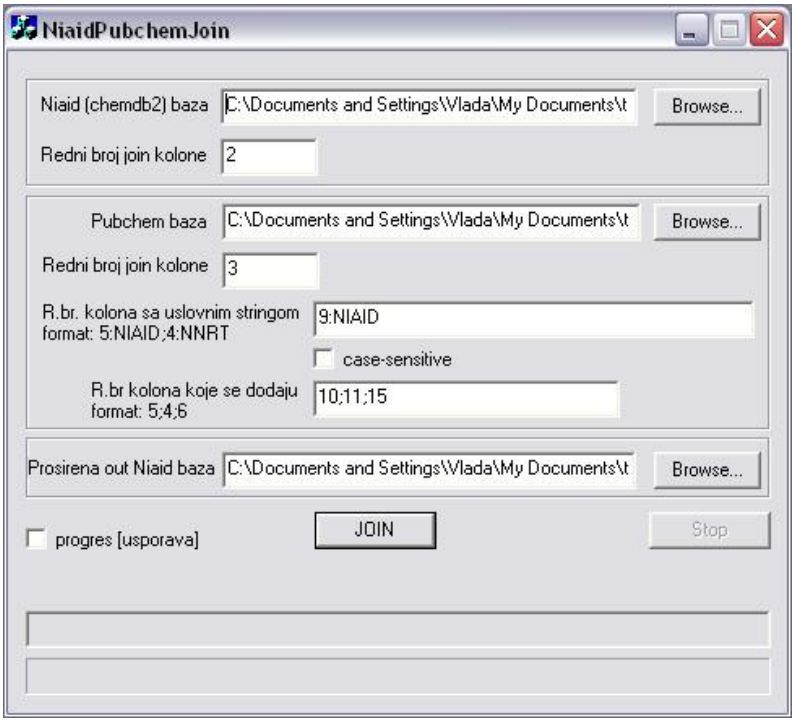
2. Parsirane su sve *.htm datoteke programom *Chemdb2Alati*
3. Za svaku *SID* vrednost, iz rezultata parsiranja, preuzete su sve *.cgi datoteke:
   http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?sid=SID&viewopt=PubChem&disopt=DisplaySDF
4. Spojene su sve *.cgi datoteke u jednu, a zatim zapakovane u *.sdf.gz
5. Primenjen je program *PubchemParser* na zapakovanu *.gz datoteku

- Lista uslova sa rednim brojevima kolona u PubChem ulaznoj bazi i niskama sa kojima se upoređuje. Format je [*N:USLOV_N; M:USLOV_M;...*]

Izlaz:

- Proširena NIAID tabela kolonama iz PubChem baze.

Program uparuje ulazne tabele po *AIDSNum* i *NIAID_ID* kolonama (slično kao JOIN upit u SQL jeziku). Ako je zadata lista uslova, rezultat se filtrira po svim uslovima iz liste, gde format uslova *N:USLOV*_N znači da niska u koloni *N* mora biti jednaka niski *USLOV_N*.



Slika 4.2.32. Prozor programa *NiaidPubchemSpoj*.

4.2.9.3. Program *FormulaKalkulator*

Svrha:

- Statistička analiza velike baze jedinjenja, po vrednostima AQVN i EIIP.
- Traženje statistički značajnog intervala i podskupa jedinjenja pogodnih za određenu biološku funkciju.

Ulaz:

- Molekularna formula nekog jedinjenja
- Baza jedinjenja sa listom molekularnih formula
- Srednja vrednost i standardna devijacija AQVN i EIIP vrednosti skupa jedinjenja za upoređivanje.

Izlaz:

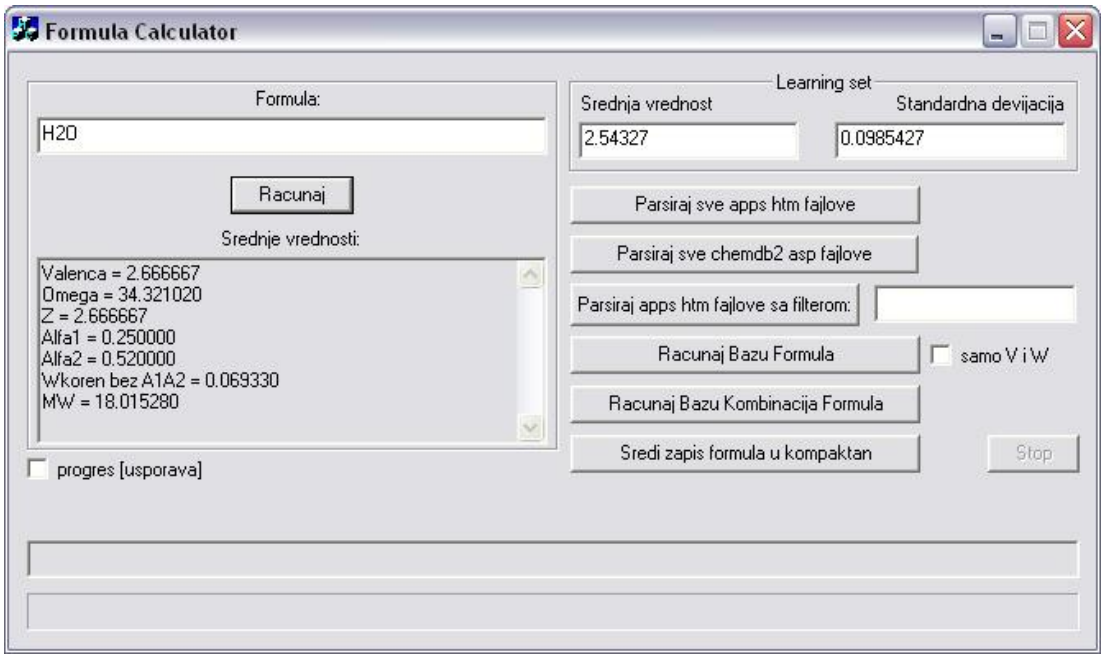
- Baza molekularnih formula sa vrednostima AQVN, EIIP.

Program parsira ulaznu molekularnu formulu svakog jedinjenja iz ulazne baze i računa vrednosti: AQVN, EIIP i molekulsku težinu (eng. *Molecular Weight, MW*). Pomoću posebnih komandi računaju se vrednosti *GausNorm* na osnovu zadatih vrednosti AQVN i EIIP skupa formula za upoređivanje. Izlazna datoteka je oblika tabele gde se za svaku formulu ispišu sva polja iz ulazne baze i izračunate vrednosti (AQVN, EIIP, MW i *GausNorm*). Za formulu sa pogrešnim zapisom ispiše se poruka o grešci u izlaznoj tabeli. U poslednja dva reda izlazne tabele dopišu se srednje vrednosti i standardno odstupanje za svaku kolonu izračunatih vrednosti.
Pomoću različitih komandi, koje imaju dodatne funkcionalnosti, obrađuju se različite vrste ulaznih baza. Komande programa su:

- *Racunaj Bazu Formula*. Pretpostavlja se da je u poslednjoj koloni ulazne datoteke zapis formule i računa vrednosti: AQVN, EIIP, MW i *GausNorm*. Ako je izabrana opcija *samo V i W*, onda se računaju samo vrednosti: AQVN i EIIP.
- *Parsiraj sve apps htm datoteke*. Parsiraju se sve *.htm datoteke preuzete sa Chemdb internet stranice, koje se nalaze u zadatom direktorijumu i pretražuju sledeća polja koja se upisuju u tabelu jedinjenja: *ChemicalName, Formula,*

Company, Classes, Anti-HIV Cellular data, Anti-HIV Enzyme data. Zatim se izračunaju i dopišu AQVN i EIIP vrednosti.

- *Parsiraj sve apps htm datoteke sa filterom* je kao prethodna komanda, ali filtrira samo ona jedinjenja koja u polju *Classes* sadrže zadatu nisku.
- *Racunaj Bazu Kombinacija Formula*. Ulazna tabela sadrži delove formula raspoređene po poljima. U svakom redu prvo polje određuje koliko kolona se posmatra. Za svaki red se generišu sve moguće formule spajanjem zadatih delova formula. Zatim se izračunaju vrednosti AQVN, EIIP, MW i *GausNorm*.
- *Sredi zapis formula u kompaktan*: sređuje ulaznu tabelu tako što spaja sve kolone u jednu kolonu koja je zapis formule.



Slika 4.2.33. Prozor programa *FormulaKalkulator*.

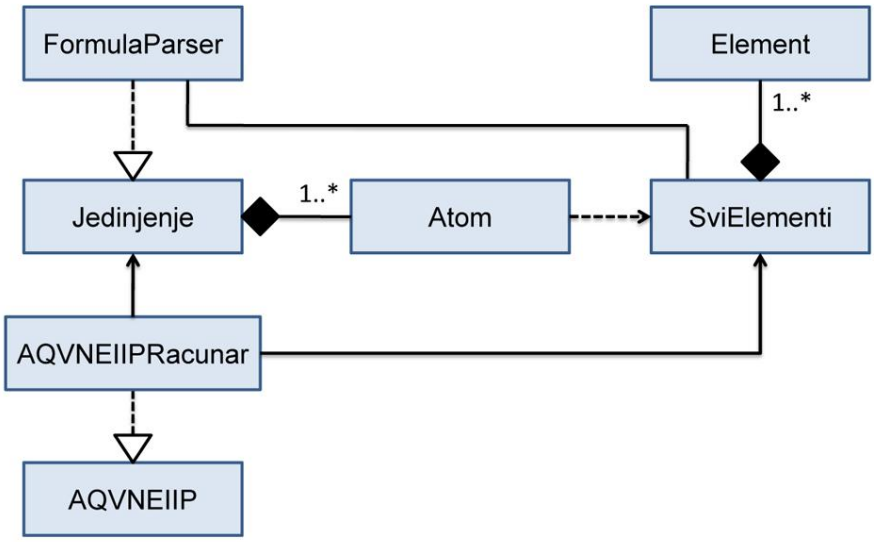
Implementacija programa *FormulaKalkulator*

Osnovne klase implementirane u programu FormulaKalkulator, pored klasa jezgra EIIP/ISM platforme, su:

- *Element* sadrži podatke o hemijskom elementu: simbol elementa, valencu, AQVN, EIIP, molekulsku težinu i atomsku masu;

- *SviElementi* skladišti niz svih hemijskih elemenata Mendaljejevog sistema u nizu objekata tipa *Element* sa izračunatim AQVN i EIIP vrednostima;
- *Atom* poseduje referencu na elemenat u objektu *SviElemeni*;
- *Jedinjenje* sadrži niz objekata *Atom* od kojih se jedinjenje sastoji;
- *FormulaParser* ima funkciju da sintaksno analizira zapis formule jedinjenja i generiše objekat *Jedinjenje*;
- *AQVNEIIPRacunar* računa AQVN i EIIP vrednosti objekta *Jedinjenje* na osnovu AQVN i EIIP vrednosti pojedinih elemenata od kojih je jedinjenje sastavljeno i generiše objekat *AQVNEIIP* koji skladišti izračunate vrednosti za datu formulu.

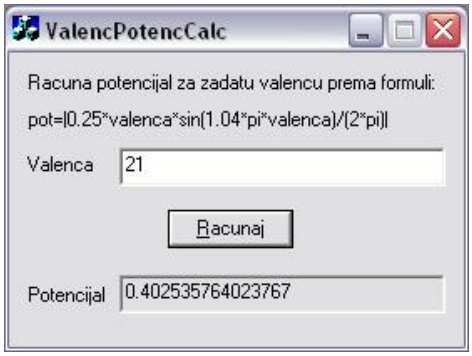
Na slici 4.2.34 je prikazan odnos osnovnih klasa u programu FormulaKalkulator.



Slika 4.2.34. Klasni dijagram programa FormulaKalkulator.

4.2.9.4. Program *ValencPotencKalk*

Program je u obliku jednostavnog dijaloga za računanje EIIP potencijala molekula za zadatu vrednost valence AQVN. Predstavlja jednostavnu verziju programa *FormulaKalkulator*.



Slika 4.2.35. Prozor programa *ValencPotencKalk*.

4.2.9.5. Program *PubchemParser*

Svrha:

- Prikupljanje SMILES zapisa svih formula iz *.txt.gz datoteka iz PubChem baze.

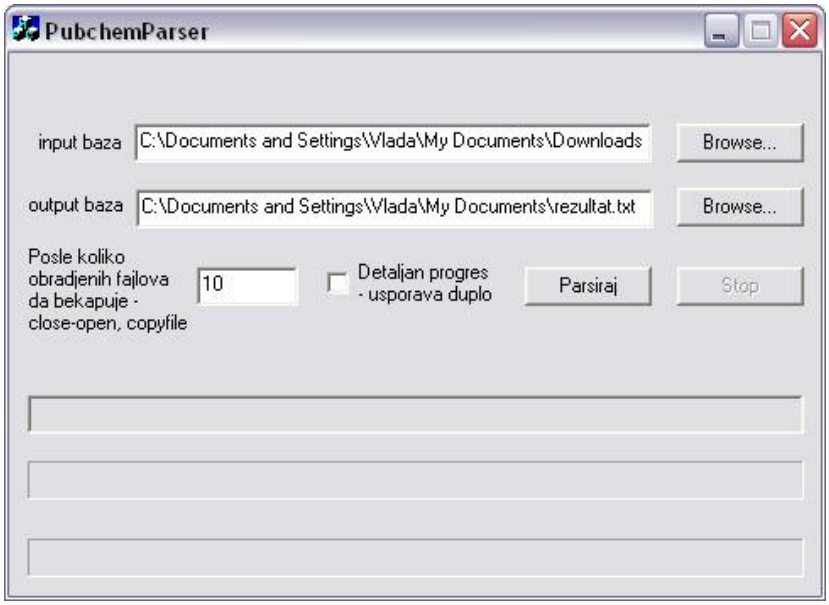
Ulaz:

- Direktorijum sa zapakovanim *.sdf datotekama (*.sdf.gz) preuzetih sa PubChem internet servisa.

Izlaz:

- Tabela hemijskih jedinjenja sa molekulskom formulom i SMILES zapisom.

Svaka datoteka iz ulaznog direktorijuma se prvo otpakuje (npr. korišćenjem programa *winrar.exe*), zatim se otpakovana *.sdf datoteka parsira i pretražuju hemijska jedinjenja sa sledećim poljima: *formula*, *openeye_name*, *cid*, *iso_smile zapis*, *komentar*, *sinonim naziva*, *regid*. Jedinjenje sa pronađenim informacijama se dodaje u tabelu rezultata. Pošto proces traje veoma dugo zbog velike količine podataka, posle nekog vremena se pravi rezervna kopija trenutnog rezultata, a status obrade se dopisuje u log datoteku.



Slika 4.2.36. Prozor programa *PubchemParser*.

4.2.9.6. Program *PubchemTxtParser*

Svrha:

- Pretraživanje PubChem baze i preformatiranje traženih informacija u tabelarni oblik.

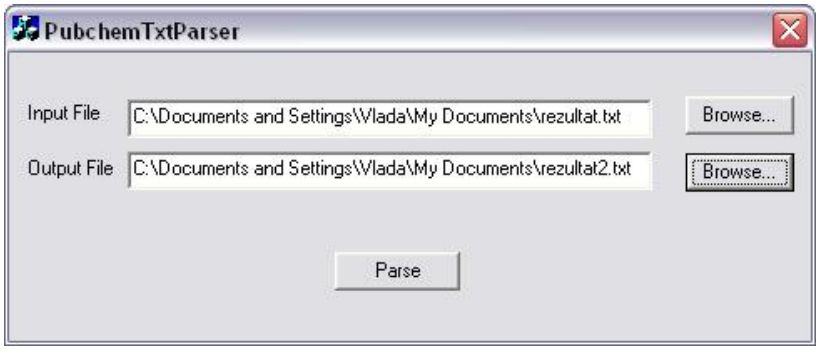
Ulaz:

- PubChem baza jedinjenja u SDF formatu (datoteka tipa *.sdf).

Izlaz:

- Tabela informacija o jedinjenjima: CID, IUPAC, MW, Formula, broj aktivnih, broj testiranih.

Program parsira ulaznu bazu PubChem jedinjenja i pretražuje informacije: *CID*, *IUPAC* naziv, molekulsku težinu (*MW*), broj testiranih i broj aktivnih u bazi BioAssays. Rezultat se zatim preformatira u tabelarni oblik.



Slika 4.2.37. Prozor programa *PubchemTxtParser*.

4.2.9.7. Program *QSARParser*

Svrha:

- Generisanje baze Anti-HIV jedinjenja, koja služi kao ulaz za dalju obradu u programima zasnovanim na *QSAR* modelu.

Ulaz:

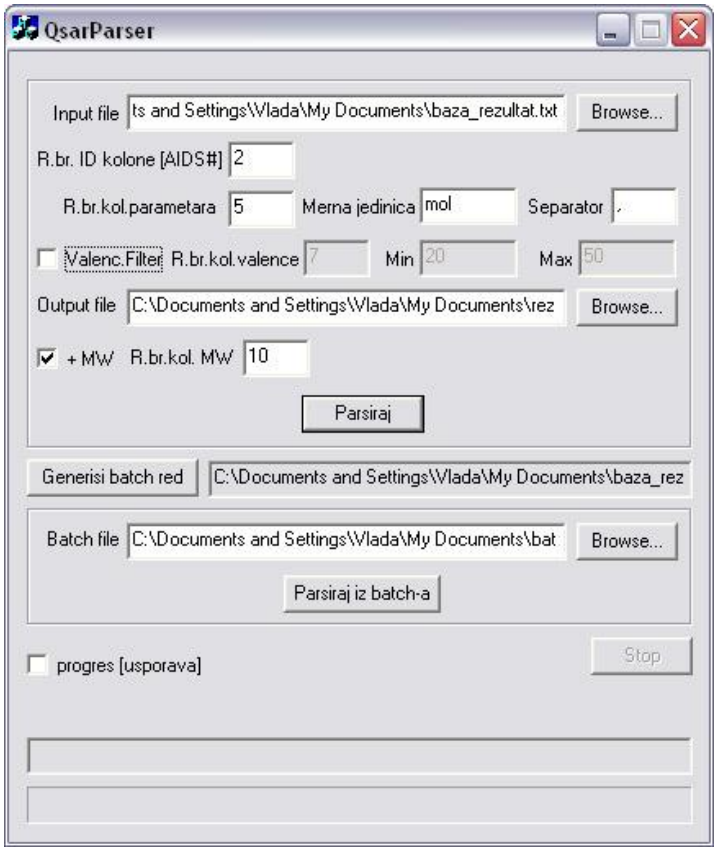
- Datoteka Anti-HIV jedinjenja u tabelarnom formatu.
- Redni brojevi kolona gde se određene informacije nalaze (*aids_num*, r.br.kolone parametara, valenca, molekulska težina)

Izlaz:

- Tabela Anti-HIV jedinjenja sa dva polja: *aids_num* i *parametar*.

Program parsira ulaznu bazu Anti-HIV jedinjenja, koja je rezultat obrade programa *ChemdbAlati*, gde za svaku vrednost parametra odvojenih separatorom tražene kolone jednog ulaznog jedinjenja, ispisuje poseban red u izlaz. Po potrebi se filtriraju jedinjenja po zadatom AQVN intervalu. Zbog velikog broja preuzetih datoteka koje se parsiraju, da se ne bi posebno za svaku ulaznu datoteku pokretao program, moguće je za ulazne parametre svake datoteke generisati komandu i svaku komandu dopisati u komandnu datoteku. Komandna datoteka se može učitati i pokrenuti

komandom *Parsiraj iz batch-a*, gde se sve komande iz komandne datoteke sekvencijalno izvršavaju.



Slika 4.2.38. Prozor programa *QSARParser*.

4.2.9.8. Program *Raspodela*

Svrha:

- Vizuelno prikazivanje gustine raspodele nekog skupa brojeva.

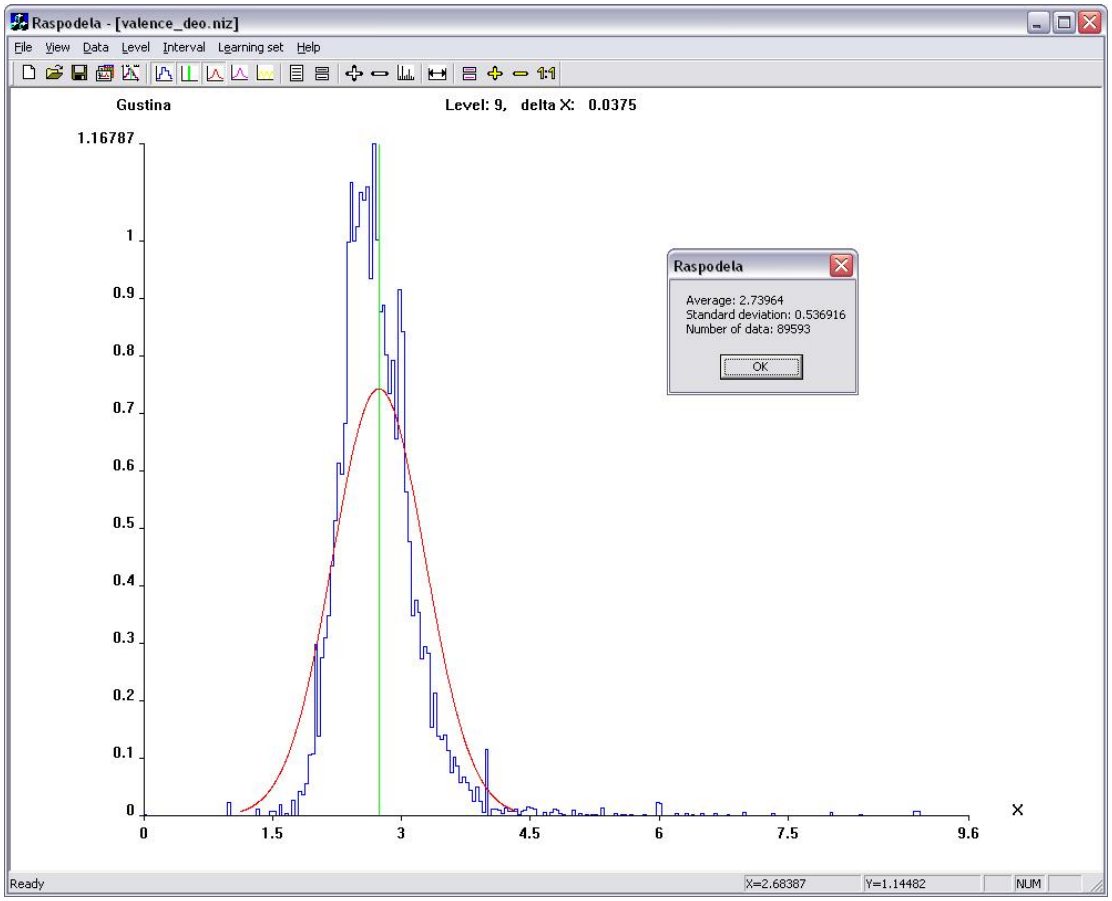
Ulaz:

- Niz brojeva snimljenih u datoteci ili direktnim unosom preko dijaloškog prozora programa, gde je svaka vrednost u posebnom redu.
- Širina malih intervala u kojima se prebrojavaju vrednosti iz ulaznog niza, odnosno rezolucija prebrojavanja.

Izlaz:

- Slika prikazane raspodele se može snimiti u datoteku tipa .bmp.
- Gustine po intervalima, gde se svaki interval zapisuje u jednom redu sa poljima: *levi, desni kraj intervala, gustina.*
- Gustine u formatu pogodnim za program *Origin*, gde se svaki interval zapisuje u jednom redu sa poljima: *centar intervala, gustina.*

Program prikazuje gustinu metodom prebrojavanja ulaznih vrednosti u malim sukcesivnim intervalima. Rezultat se prikazuje grafikom, gde je *x* osa vezana za ulazne vrednosti, a *y* osa za gustinu. Tačke na grafiku su definisane ivicama intervala (*x* koordinata) i vrednost gustine u tom intervalu (*y* koordinata). Grafik se iscrtava plavom bojom. Površina grafika uvek ima vrednost 1.



Slika 4.2.39. Osnovni prozor programa *Raspodela*.

Na grafiku se može prikazati i srednja vrednost ulaznih vrednosti kao vertikalna linija sa *x* koordinatom srednje vrednosti (zelenom bojom). Pored grafika gustine, za upoređivanje se može iscrtati i Gausova kriva (crvenom bojom) definisana sa srednjom vrednosti i standardnom devijacijom ulaznog niza. Prelaskom mišem iznad grafika u donjoj desnoj statusnoj traci se ispisuju koordinate na grafiku na poziciji miša. Rezolucija tj. širina intervala se može uneti ručno ili se interval može posebnim komandama automatski duplirati ili prepoloviti.

4.2.9.9. Program *Raspodela2D*

Svrha:

- Vizuelno prikazivanje gustine raspodele skupa dvodimenzionalnih brojeva.

Ulaz:

- Niz brojeva snimljenih u datoteci ili direktnim unosom preko dijaloškog prozora programa, u formatu takvom da je svaki dvodimenzionalni broj zapisan u posebnom redu pri čemu su prva i druga koordinata razdvojene *tab* simbolom.
- Širina i visina malog dvodimenzionalnog intervala u kojem se prebrojavaju vrednosti iz ulaznog niza, odnosno rezolucija prebrojavanja.

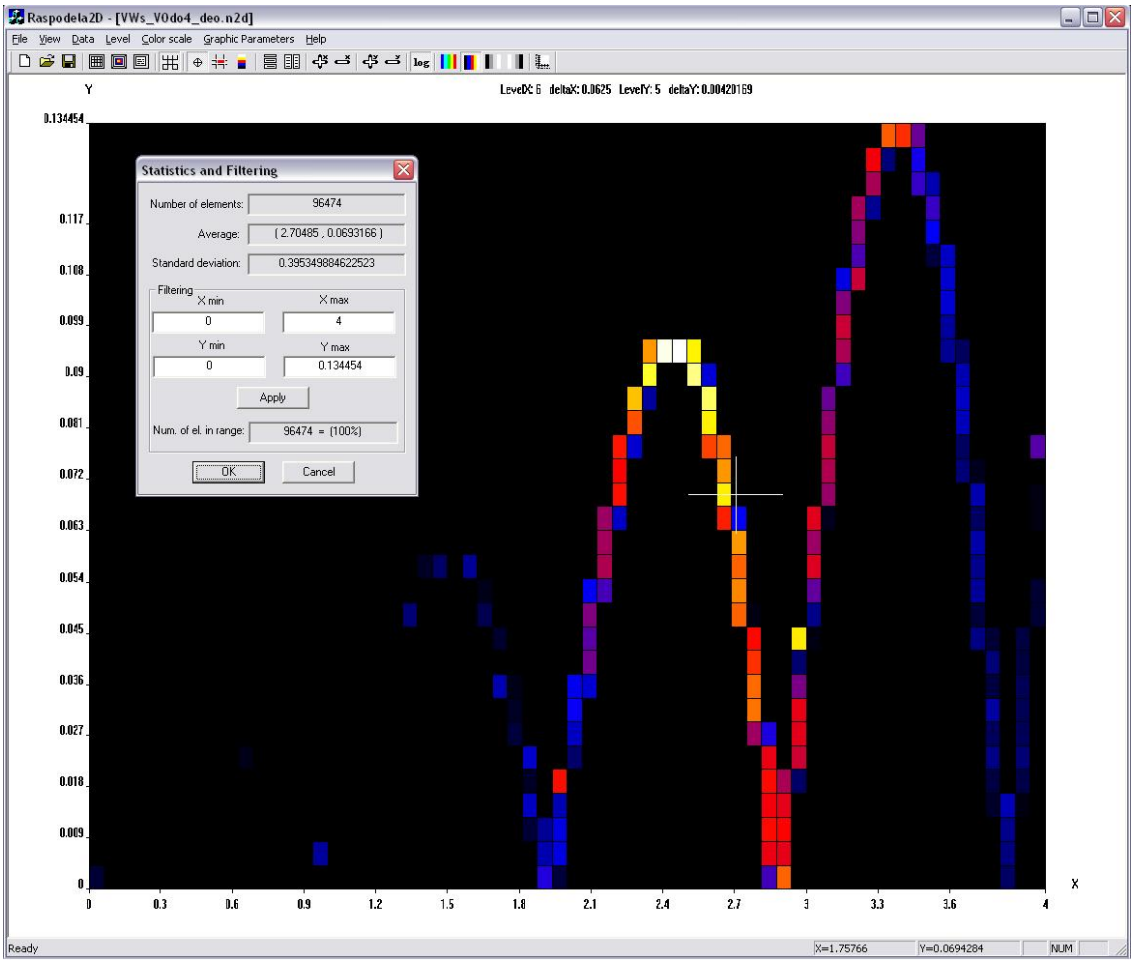
Izlaz:

- Slika prikazane raspodele, koja se može snimiti u datoteku tipa *.bmp.
- Gustine po intervalima, gde se svaki interval zapisuje u jednom redu sa poljima: *levi, desni, donji, gornji kraj intervala, gustina*.
- Gustine u formatu pogodnim za program *Origin*, gde se svaki interval zapisuje u jednom redu sa poljima: *x, y centar intervala, gustina.*
- Trodimenzionalni model (*x* i *y* osa predstavljaju ulazne vrednosti, *z* osa predstavlja vrednost gustine) predstavljen mrežom trouglova: jedna ivica trougla je ivica regiona (pravougaonika) sa *z*=0, a suprotno teme ima *x,y* koordinate centra regiona i *z* koordinatu jednaku vrednosti gustine tog regiona.
  Format zapisa svakog trougla je predstavljen temenima: *Ax, Ay, Az, Bx, By, Bz, Cx, Cy, Cz.*

Ulazni dvodimenzionalni brojevi se posmatraju u X-Y ravni. Površina se podeli u mrežu pravougaonika (dvodimenzionalnih intervala) zadate rezolucije i broje se ulazne vrednosti u svakom pravougaoniku. Rezultat se prikazuje dvodimenzionalnom mrežom gde su x i y ose vezane za ulazne vrednosti, a boja svakog regiona predstavlja gustinu. Postoje 4 šablona skala boja, uz izbor linearne ili logaritamske skale. Na grafiku se može prikazati i srednja vrednost ulaznih vrednosti.

Prelaskom mišem iznad grafika u donjoj desnoj statusnoj traci se ispisuju koordinate na grafiku na poziciji miša. Rezolucija, tj. širina i visina intervala se mogu uneti ručno ili se interval može posebnim komandama automatski duplirati ili prepoloviti, posebno po visini i po širini.

U dijalogu za statistiku i filtriranje se ispisuju srednja vrednost i standardna devijacija, uz mogućnost izračunavanja procenta prisutnih vrednosti u zadatom intervalu.



Slika 4.2.40. Osnovni prozor programa *Raspodela2D*.

4.2.10. Modul za obradu zapisa sekvenci

Modul za obradu zapisa sekvenci sadrži razne alate za ručnu i automatsku obradu tekstualnog zapisa sekvenci, pretraživanje iz baza sekvenci, automatsku transformaciju zapisa sekvenci, generisanje baze mutiranih sekvenci na osnovu liste mutacija i prevođenje DNK zapisa u proteinske sekvence. Modul se sastoji iz programa: SekEditor, SekuFasta, ProteinBazaSec, DNKuProtein, FastaMutGen, FastaFilter i GenomNetFilter.

4.2.10.1. Program *SekEditor*

Svrha:

- Vizuelno okruženje za obradu tekstualnog zapisa proteinskih sekvenci.

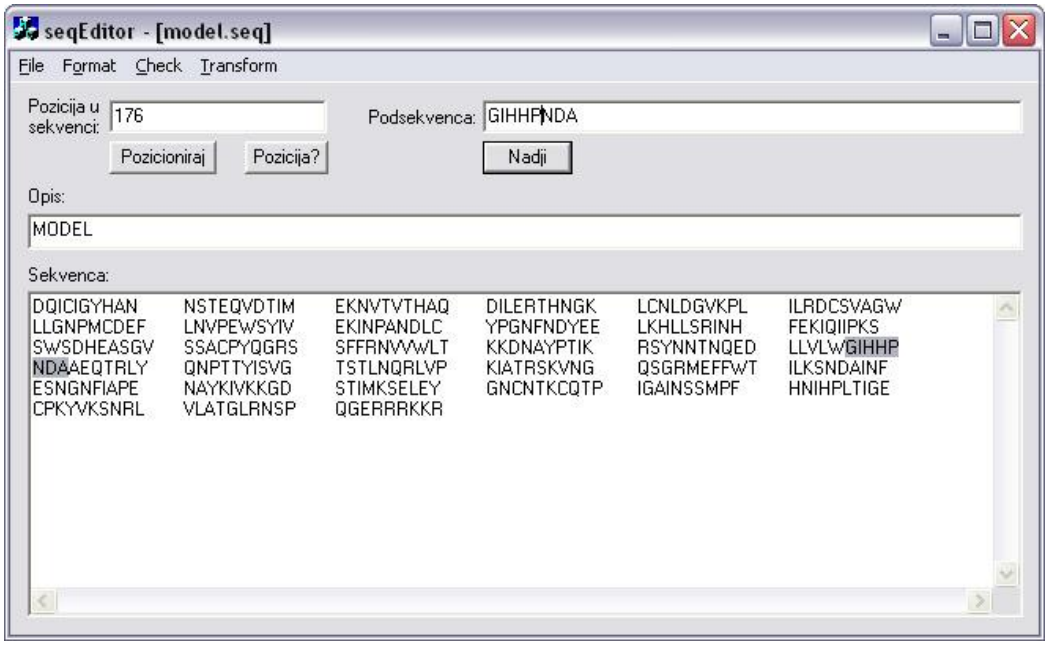
Ulaz:

- Datoteka tipa *.seq
- Direktan unos sekvence u tekstualno polje
- Datoteka u FASTA, GenBank ili Swissprot formatu.

Izlaz:

- Datoteka u SEQ formatu

Osim otvaranja i snimanja datoteka u *.seq ekstenziji, odnosno SEQ formata, moguće je uneti zapis proteina iz datoteka u FASTA, GenBank ili Swissprot formatu. Pored osnovnog editovanja zapisa sekvenci i opisa, postoje i dodatne opcije: formatiranje zapisa sa *tab* simbolima od po 10 kiselina u grupi ili 80 u jednom redu; pretvaranje malih u velika slova; brisanje karaktera koji nisu slova i prazni znaci; provera ispravnosti jednoslovnog ili troslovnog zapisa proteina; brojanje kiselina u sekvenci; transformacija jednoslovnog zapisa u troslovni i obrnuto; traženje podsekvence; detekcija selektovane pozicije; pozicioniranje na traženu poziciju u sekvenci, uz ignorisanje praznih znakova (belina) u zapisu sekvence.



Slika 4.2.41. Osnovni prozor programa *SekEditor*.

4.2.10.2. Program *SekuFasta*

Program formira FASTA datoteku od svih sekvenci iz *.seq datoteka, koje se nalaze u zadatom direktorijumu.

4.2.10.3. Program *ProteinBazaSec*

Svrha:

- Generisanje unificiranih sekvenci iz baze proteina, odnosno izbacivanje nepotpunih sekvenci i odsecanje nepotrebnih početaka i krajeva sekvenci. Time se uravnotežuju dužine sekvenci i izbacuju one sa velikim odstupanjem.

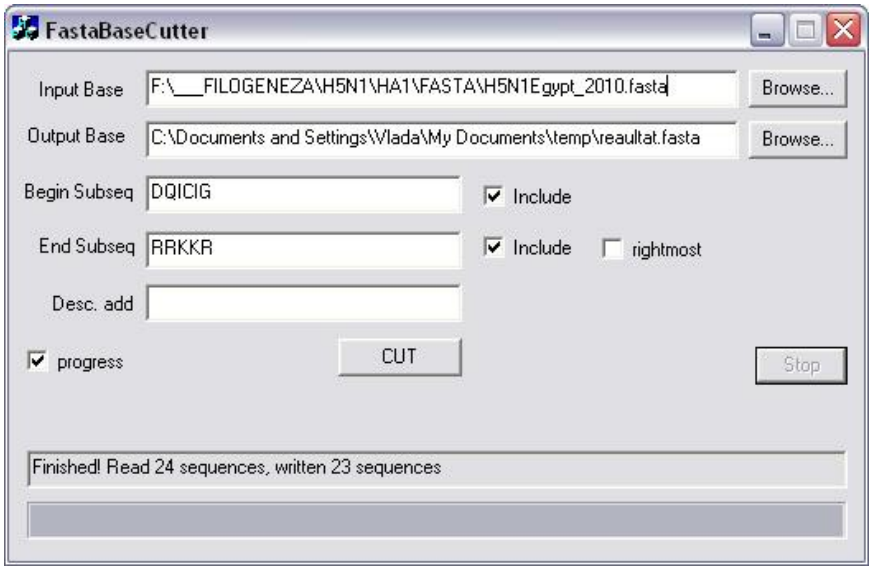
Ulaz:

- Datoteka proteina u FASTA ili SwissProt formatu
- Početak podsekvence - niz proteina kojom počinje tražena podsekvenca
- Kraj podsekvence - niz proteina kojom se završava podsekvenca
- Dodatan opis

Izlaz:

- Datoteka proteina u FASTA ili SwissProt formatu
- Broj pronađenih podsekvenci i ukupan broj sekvenci u ulaznoj bazi

Iz ulazne baze proteina za svaku sekvencu se traži podsekvenca koja počinje zadatim početkom, a završava se sa zadatim krajem. Odsecaju se delovi proteina pre i posle pronađene podsekvence. Može se izabrati da li će se zadate niske naći u rezultujućoj sekvenci ili ne. Ako nije pronađena takva podsekvenca, sekvenca ne ulazi u rezultat. Za pretragu zadatog početka traži se najdešnja podniska u sekvenci, a za zadati kraj može se izabrati krajnje levi ili krajnje desni, tako da podsekvenca bude najkraća ili najduža moguća.



Slika 4.2.42. Prozor programa *ProteinBazaSec*.

4.2.10.4. Program *DNKuProtein*

Svrha:

- Prevodi DNK sekvence u proteinske sekvence

Ulaz:

- Baza DNK sekvenci u mogućim formatima:

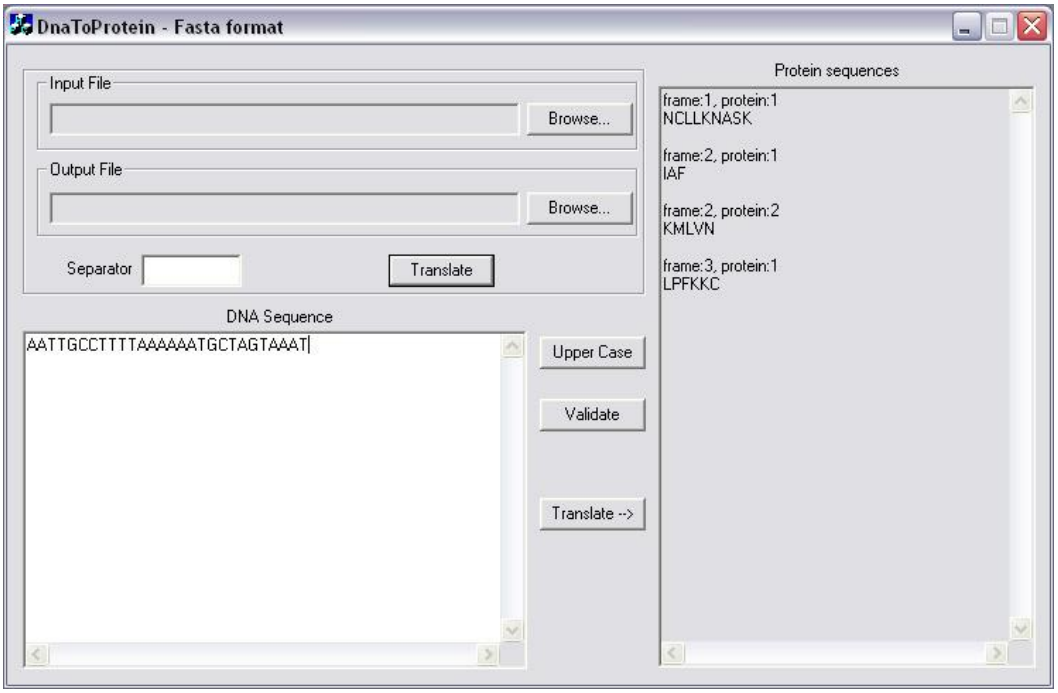
- Sekvence iz svih *.seq datoteka iz određenog direktorijuma
- Datoteka u FASTA, SwissProt ili GenBank formatu

- Separator između svih mogućih prevedenih proteinskih sekvenci jedne DNK sekvence, ili ako nije zadat onda se sve zapisuju u izlazu kao posebne sekvence
- Direktan unos DNK sekvence u tekstualno polje

Izlaz:

- Baza proteinskih sekvenci u formatima FASTA, SwissProt ili GenBank
- Tekstualno polje u slučaju direktnog prevođenja

Za ulazne DNK sekvence generišu se sve moguće proteinske sekvence i to na sledeći način. Za svaku DNK sekvencu moguće je početi prevođenje sa prve, druge ili treće pozicije (tri moguća frejma), a posle svakog stop kodona počinje novo prevođenje. Na taj način, za svaku ulaznu DNK sekvencu postoji ukupno 3*(*broj_stop_kodona*+1) mogućih proteinskih sekvenci. Za prevođenje DNK sekvence u protein se koristi standardni genetski kôd (tabela 4.2.2).



Slika 4.2.43. Prozor programa *DNKuProtein*.

U slučaju direktnog tekstualnog prevođenja, moguće komande su:

Validate: proverava da li je ispravan zapis DNK sekvence

Upper Case: menja mala u velika slova u DNK zapisu

Translate: prevodi DNK sekvencu u proteinske sekvence odvojene praznim redom, sa informacijom o početnoj poziciji (frejmu) i rednom broju sekvence.

Tabela 4.2.2. Standardni genetski kod.

Aminokiselina	**Jednoslovni zapis**	**DNK kodon**
Izoleucin	I	ATT, ATC, ATA
Leucin	L	CTT, CTC, CTA, CTG, TTA, TTG
Valin	V	GTT, GTC, GTA, GTG
Fenilalanin	F	TTT, TTC
Metionin	M	ATG
Cistein	C	TGT, TGC
Alanin	A	GCT, GCC, GCA, GCG
Glicin	G	GGT, GGC, GGA, GGG
Prolin	P	CCT, CCC, CCA, CCG
Treonin	T	ACT, ACC, ACA, ACG
Serin	S	TCT, TCC, TCA, TCG, AGT, AGC
Tirozin	Y	TAT, TAC
Triptofan	W	TGG
Glutamin	Q	CAA, CAG
Asparagin	N	AAT, AAC
Histidin	H	CAT, CAC
Glutaminska kiselina	E	GAA, GAG
Asparaginska kiselina	D	GAT, GAC
Lizin	K	AAA, AAG
Arginin	R	CGT, CGC, CGA, CGG, AGA, AGG
Stop kodoni	Stop	TAA, TAG, TGA
Lizin	K	AAA, AAG
Arginin	R	CGT, CGC, CGA, CGG, AGA, AGG
Stop kodoni	Stop	TAA, TAG, TGA

4.2.10.5. Program *FastaMutGen*

Svrha:

- Generiše kompletne zapise svih mogućih mutiranih sekvenci na osnovu liste mutacija i zapisa osnovne sekvence.

Ulaz:

- Baza sekvenci u FASTA formatu
- Lista mutacija sledećeg formata: svaki red sadrži jedno ili nekoliko pravila supstitucija odvojenih zarezom. Supstitucija je oblika *XNY* ili *NY*, gde je *N* pozicija kiseline u sekvenci, *Y* nova mutirana kiselina, *X* stara kiselina. *Y* može biti simbol „~“ koji označava terminaciju (stop kodon) ili simbol „–“ koji označava deleciju.

Izlaz:

- Baza sekvenci u FASTA formatu

Program mutira svaku sekvencu iz ulazne baze redom, koristeći svako pravilo iz liste pravila, tako što sva pravila iz jednog reda liste upotrebi zajedno. U izlaznoj bazi ima ukupno $N(K+1)$ sekvenci, gde su: N broj sekvenci u ulaznoj bazi, K broj redova u listi pravila, +1 zato što u izlazu doda originalnu sekvencu nepromenjenu.

4.2.10.6. Program *FastaFilter*

Svrha:

- Filtriranje ulazne baze po poljima u opisu sekvenci. Primer je izdvajanje samo humanih proteina unosom teksta HUMAN u polje *EntryName*.

Ulaz:

- Datoteka proteina u FASTA formatu.
- Polja za pretragu:
    - *EntryName*: polje iz opisa sekvence

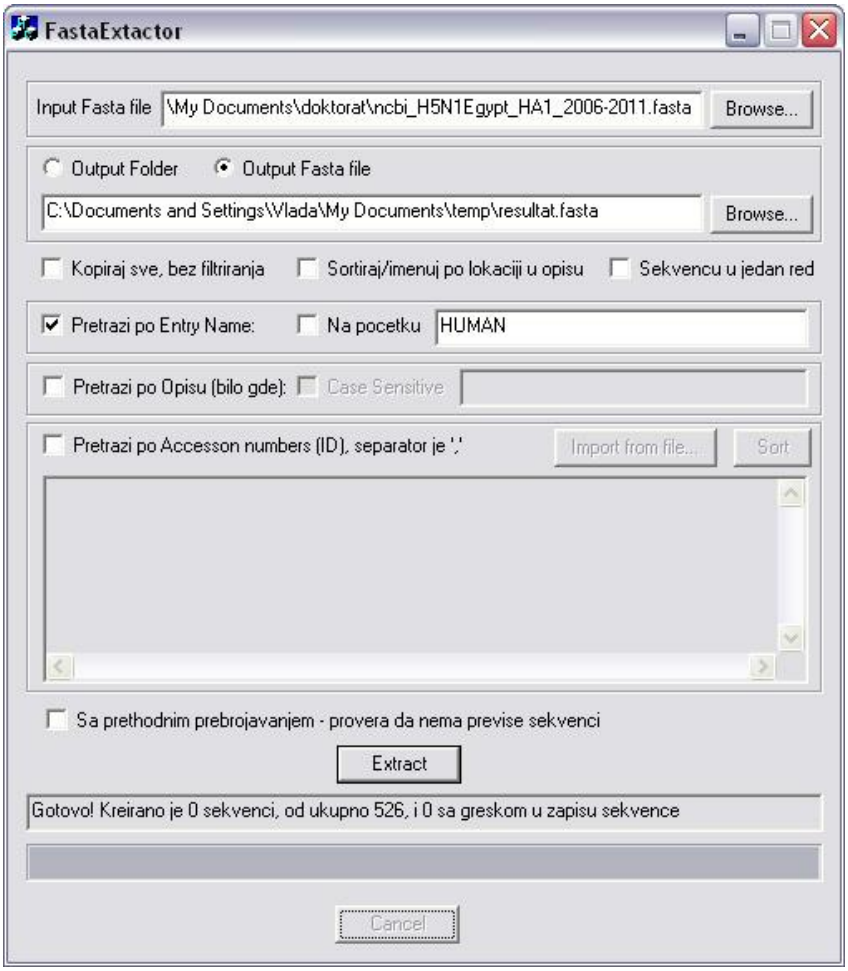
- Opis sekvence iza *EntryName* dela
- Lista ID vrednosti, koja se može uneti: (i) direktno u tekstualno polje, gde su ID vrednosti razdvojeni simbolom „ , “ ili (ii) iz tekstualne datoteke, gde je u svakom redu jedna ID vrednost.

Izlaz:

- Baza proteina u FASTA formatu
- Datoteke tipa *.seq, gde svaka datoteka dobije naziv prema podacima iz opisa sekvence: *entryName* [*ID*]_*num*.seq.
- Informaciju o broju pronađenih sekvenci i ukupnom broju sekvenci u ulaznoj bazi.

Program pretražuje iz ulazne baze proteine koje zadovoljavaju jedan od tri zadata uslova pretrage (ne moraju biti zadata sva tri uslova). Opis sekvence u FASTA formatu je oblika *>ID|entryName opis*. Prvi uslov pretrage je po *EntryName* delu iz opisa sekvence, gde se može uključiti opcija *na početku* tako da traži samo one sekvence gde se zadata niska nalazi kao podniska na početku polja *EntryName*, inače bilo gde. Drugi uslov pretrage je po opisu, gde se zadata niska traži kao podniska u *opis* delu sekvence. Treći je po listi ID vrednosti, gde *ID* iz opisa sekvence mora biti identičan nekom iz zadate liste. Ako je izabrana opcija *sortiraj/imenuj po lokaciji u opisu*, u izlaznu bazu se ispisuju nađene sekvence sortirane po lokacijama, gde je *lokacija* deo u opisu sekvence između uglastih zagrada „..[*lokacija*].." , a u slučaju *.seq datoteka, nazivu datoteke se na početak doda *lokacija*.

Postoji opcija u programu koja omogućava kopiranje svih sekvenci bez uslova, a time i prevod zapisa sekvenci iz jednog formata u drugi.



Slika 4.2.44. Prozor programa *FastaFilter*.

4.2.10.7. Program *GenomeNetFilter*

Svrha:

- Kreiranje FASTA datoteke sa sekvencama preuzetih sa GenomeNet internet stranice spremne za dalju ISM analizu. Kasnije se programom *FastaFilter* mogu iz FASTA baze isfiltrirati sekvence određenih organizama pretraživanjem po polju *EntryName*.

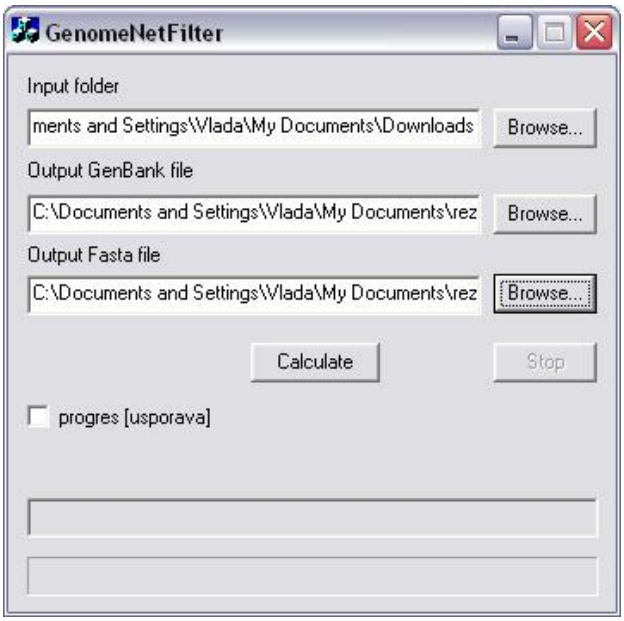
Ulaz:

- Preuzete *.htm datoteke sa GenomeNet internet stranice.

Izlaz:

- Baza svih sekvenci u GenBank ili FASTA formatu

Program parsira svaku *.htm datoteku u kojoj je jedna sekvenca u GenBank formatu i doda je u zajedničku datoteku GenBank formata koja sadrži sve sekvence. Zatim, od te zajedničke datoteke kreira datoteku u FASTA formatu sa istim sekvencama gde za opis uzima *ACCESSION* i *DEFINITION* polja iz GenBank datoteke, a opis je oblika: >*ACCESSION*|*ACCESSION DEFINITION*.

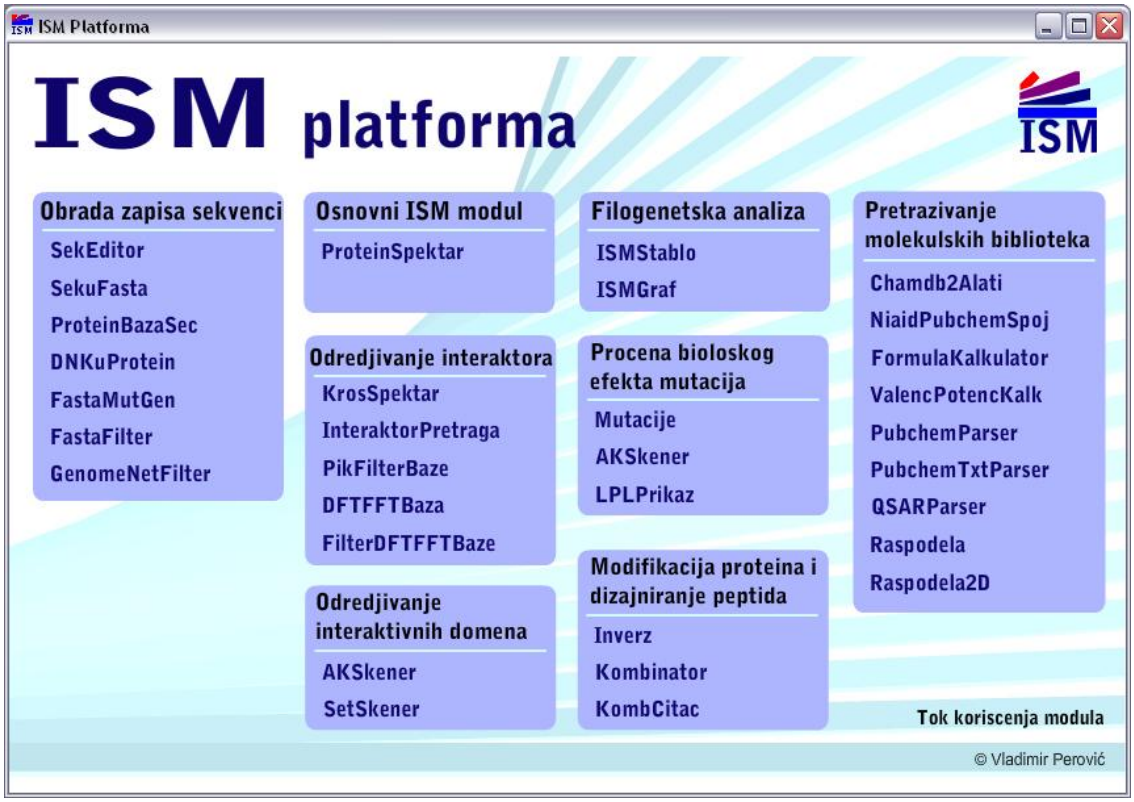


Slika 4.2.45. Prozor programa *GenomeNetFilter*.

4.2.11. Program za objedinjavanje modula u platformu

Razvoj pojedinačnih programa zasnovanih na ISM metodi je trajao postepeno. Kako su se javljale nove potrebe i ideje, tako su razvijani posebni programi, a funkcionalno slični su spajani u zasebne module. Na kraju su svi moduli i programi organizovani i objedinjeni u formi platforme.

Razvijena je centralna aplikacija *ISMPlatforma* sa svrhom pozivanja posebnih programa platforme. Grafički korisnički interfejs omogućava vizuelni pregled modula i praćenje toka korišćenja modula pogodnog za analizu dalekodosežnih međumolekulskih interakcija i dizajn lekova zasnovan na ISM metodi.



Slika 4.2.46. Prozor osnovne aplikacije ISM paketa, koji objedinjuje sve module sa pripadajućim programima.

4.3. Primene EIIP/ISM platforme

EIIP/ISM platforma je uspešno primenjena u rešavanju sledećih bioloških problema: određivanje terapeutskih i dijagnostičkih targeta, procena biološkog uticaja mutacija [222], filogenetska analiza, selekcija terapeutskih malih molekula i predikcija odgovora na kombinovanu terapiju hroničnih HCV bolesnika [223].

4.3.1. Određivanje terapeutskih i dijagnostičkih targeta

Biološki ili terapeutski target u farmaceutskim istraživanjima predstavlja protein čija se aktivnost može modifikovati lekom kojim se postiže željeni terapeutski efekat. Identifikacija terapeutskog targeta je prvi korak u procesu otkrivanja novih lekova. Terapeutski target može biti učesnik signalnog puta povezanog sa nastankom bolesti ili može sadržati mutacije koje su povezane sa nastankom bolesti.

Za određivanje terapeutskih i dijagnostičkih targeta korišćeni su programi EIIP/ISM platforme: SeqEditor, ProteinBazaSec i FastaMutGen (Modul za obradu zapisa sekvenci); ProteinSpektar (Osnovni ISM modul); KrosSpektar (Modul za određivanje interaktora); AKSkener i SetSkener (Modul za određivanje interaktivnih domena); Mutacije i AKSkener (Modul za procenu biološkog efekta mutacija).

4.3.1.1. Identifikacija strukturnih domena hemaglutinina i polimorfizama koji utiču na modulaciju interakcije svinjskog gripa H1N1 sa humanim receptorom

Novi A/H1N1 influenca virus koji se pojavio u Severnoj Americi je blisko povezan sa severno američkim H1N1/N2 svinjskim virusima. Sve do početka 2009. godine, severno američki svinjski H1N1/N2 virusi su samo sporadično inficirali ljude. A/H1N1virus je 2009. godine stekao sposobnost efikasne transmisije sa čoveka na čoveka. U ovom radu su analizirane osobine hemaglutinina (HA) A/H1N1 odgovorne za virus/receptor interakciju.

Rezultati ove studije doprinose (i) boljem razumevanju porekla novog influenca A/H1N1 virusa, (ii) praćenju molekularne evolucije virusa, (iii) predikciji *hot-spot* mutacija (mutacija na funkcionalno značajnim pozicijama) koje favorizuju interakciju

sa humanim receptorom, (iv) identifikaciju terapeutskih, dijagnostičkih targeta i vakcinskih targeta za prevenciju i tretman A/H1N1 infekcije.

Rezultati su objavljeni u sledećem radu [224]:
Veljkovic, V, Niman HL, Glisic S, Veljkovic N, ***Perovic V****, Muller CP. Identification of hemagglutinin structural domain and polymorphisms which may modulate swine H1N1 interactions with human receptor. BMC Struct Biol. 2009 Sep 9(1), 62. doi: 10.1186/1472-6807-9-62.*

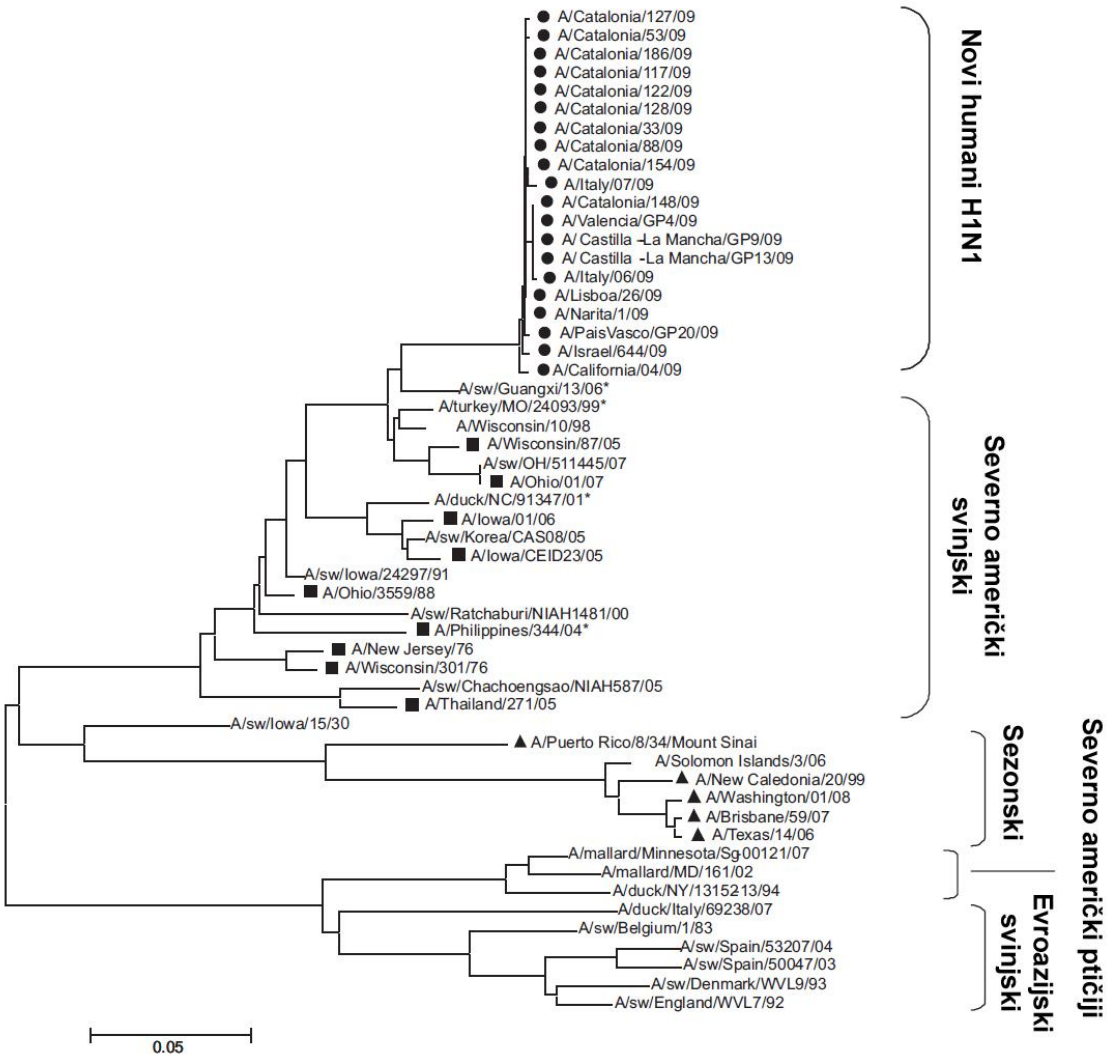
Filogenetskom analizom je pokazano da je HA A/H1N1 u najbližoj vezi sa trostrukim reasortantnim svinjskim influenca virusima H1N1 i H1N2 (segmenti RNK ovih virusa su različitog porekla, iz virusa influence koji su inficirali tri vrste domaćina: ljude, svinje i ptice) izolovanim u Severnoj Americi posle 2000. godine (slika 4.3.1.1), što ukazuje da A/H1N1 potiče od ovog virusnog klastera. Zatim su severno američki HA H1N1 i H1N2 analizirani ISM metodom i upoređeni sa A/H1N1 izolatima. Na slici 4.3.1.2. je prikazan tipičan IS profil virusa A/H1N1 (slike 4.3.1.2d,e,f), kao i konsenzus spektara virusa svake od navedenih grupa.

Krosspektar svinjskih izolata HA1 H1N2/H1N1 ima karakteristični dominantni pik na frekvenci F(0.055) a karakterističan dominantan pik za A/H1N1izolate je na frekvenci F(0.295) (slike 4.3.1.2a,c).

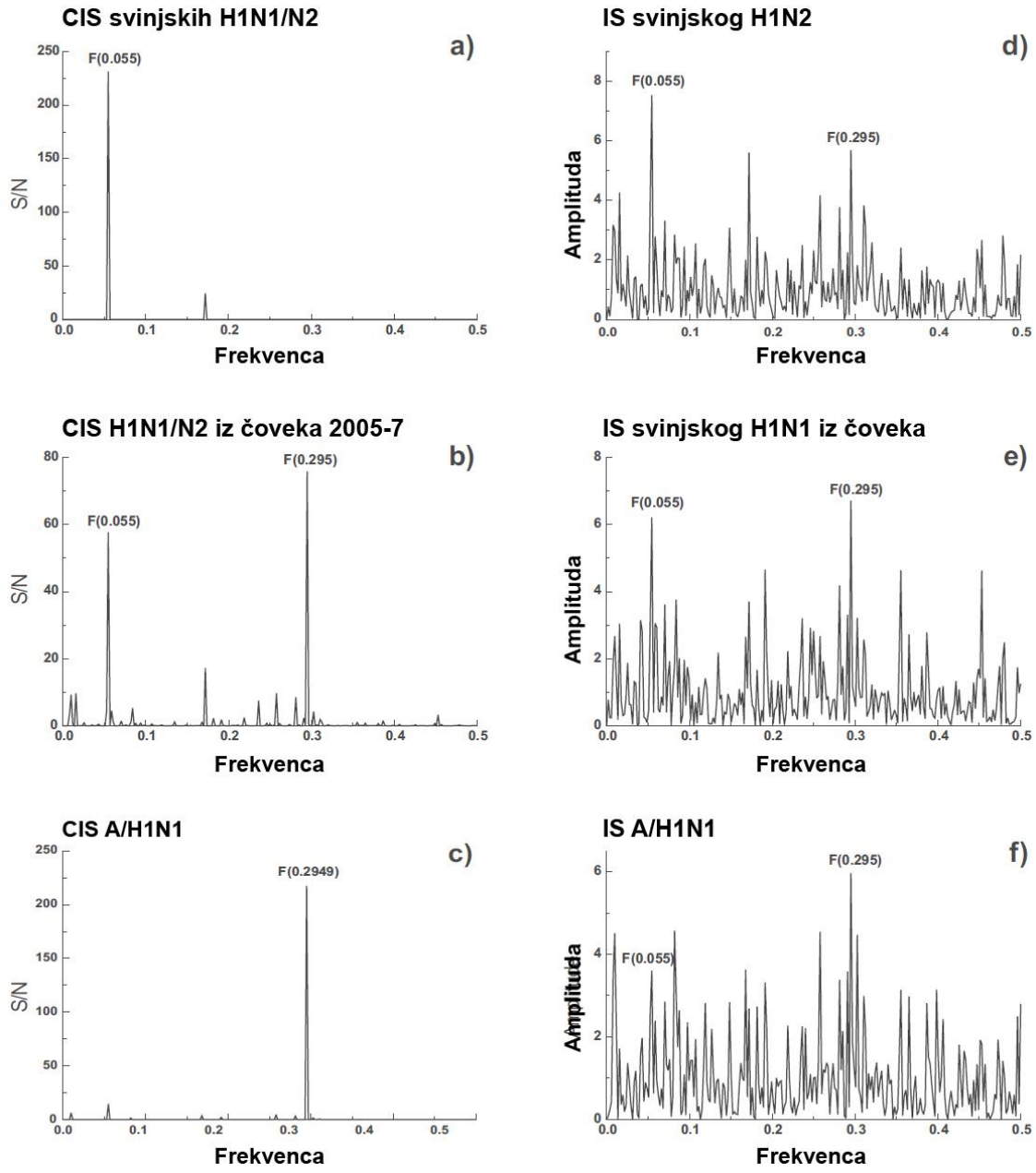
Prema ISM konceptu, to ukazuje da HA1 navedene dve grupe virusa ostvaruju drugačiji oblik interakcije sa receptorom. Na osnovu virusnog tropizma može se zaključiti da veća vrednost amplitude na F(0.055) odgovara većem afinitetu za proteine svinja, dok F(0.295) predstavlja veća afinitet za interakciju sa humanim proteinima.

CIS svinjskih HA1 izolovanih iz ljudi u SAD pre 2008. sadrži karakteristične pikove na obe frekvence F(0.055) i F(0.295) (slika 4.3.1.2b), što sugeriše da su navedeni virusi, koji su sporadično inficirali ljude, mogli da ostvare interakciju sa oba tipa receptora. Iz svega izloženog bi se moglo zaključiti da sve tri navedene grupe virusa imaju sposobnost da interreaguju i sa svinjskim i sa humanim proteinima. Navedeni rezultati takođe daju dodatni snažan dokaz da su humani izolati svinjskih virusa H1N2/H1N1, koji su cirkulisali pre 2008. u SAD, bili prekursori A/H1N1.

Veoma je važno utvrditi promene kod H1N1/N2 virusa koje su dovele do povećanog afiniteta za humani receptor. Poređenjem HA1 H1N1/N2 sekvenci svinjskih izolata sa ranim meksičkim A/H1N1 izolatom A/Mexico/4115/2009 je ustanovljeno 14 aminokiselinskih supstitucija koje su visoko specifične za A/H1N1 viruse (tabela 4.3.1.1).



Slika 4.3.1.1. Filogenetska analiza HA proteina virusa H1N1 i H1N2 koji su inficirali ljude. Humani izolati svinjskih virusa H1 između 1976. i 2007. (crni kvadrat), ptičji (H1), sezonski humani H1N1 (crni trougao), reprezentativni predstavnici novog A/H1N1 i H1N2 (zvezdica).



Slika 4.3.1.2. ISM analiza HA1 proteina. (a) CIS svinjskih H1N1 i H1N2 influenca virusa izolovanih u Severnoj Americi između 1931 i 2008, (b) CIS humanih izolata svinjskih virusa H1N1 u SAD tokom 2005-2007, A/Iowa/01/2006; A/Wisconsin/87/2005; A/Ohio/01/2007; A/Ohio/02/2007, (c) CIS A/H1N1 virusa predstavljenih na slici 4.3.1.4. (d) IS reprezentativnog svinjskog H1N2 HA1 (A/swine/Minnesota/1192/2001), (e) IS HA1humanog izolata svinjskog H1N1 (A/Iowa/01/2006), (f) IS A/H1N1 HA1 (A/Castilla-La Mancha/GP13/2009).

Tabela 4.3.1.1. Uticaj polimorfizama u HA1 na amplitude odgovarajućih IS frekvenci F(0.055) i F(0.295).

Mutacija	**ΔA [F(0.055)](%)**	**ΔA [F(0.295)](%)**	**Afinitet za humani receptor**
R36K*	-7.5	+5.0	+
L61I	0	0	0
F71S	-1.2	+0.4	+
N97D	-6.6	-9.1	±
T128S	-1.4	+1.2	+
R130K	-5.1	-5.0	±
R146K	-7.5	-2.9	±
N168D*	+4.9	-4.8	-
T216I*	-5.7	+10.5	+
A224E	0	0	0
S271P*	-4.3	+2.7	+
V298I	0	0	0
E302K	-3.9	+1.7	+
M314L	-3.6	+6.5	+

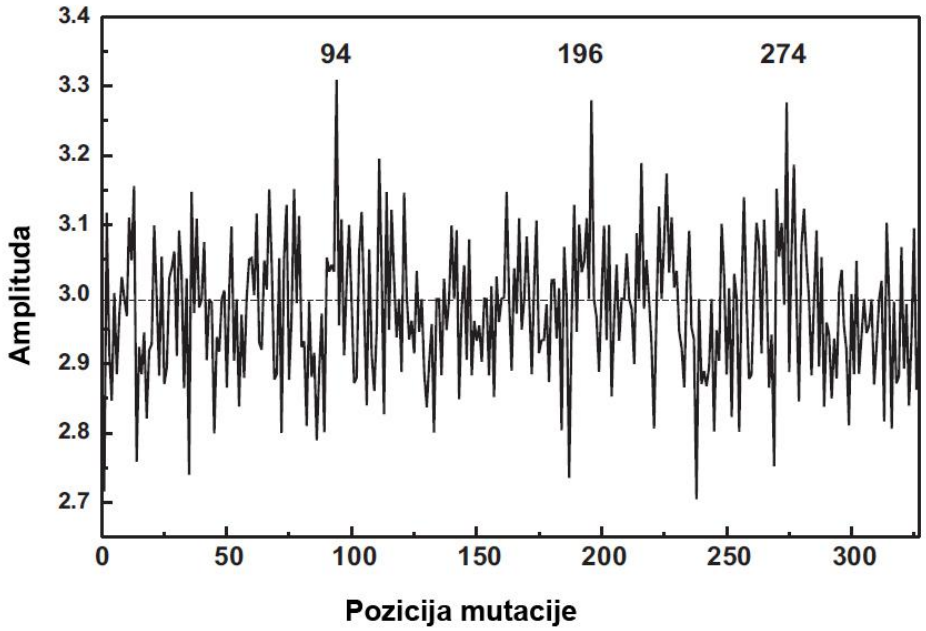
* mutacije prisutne u svinjskom H1N1/N2 soju koje inficiraju čoveka su označene zvezdicom

Zatim je ispitivano koja je od ovih mutacija pojedinačno (ili njihova kombinacija) najvažnija za promenu povećanje afiniteta za humane receptore. Kao što je već pokazano interakcija između H1N1/N2 i svinjskih proteina se karakteriše frekvencom F(0.055), a interakcija između A/H1N1 i humanih proteina na frekvenci F(0.295). Prema ISM konceptu [225, 226], mutacije u HA1 koje povećavaju amplitudu na F(0.295) i smanjuju amplitudu na F(0.055) bi mogle da doprinesu promeni virusnog tropizma od svinjskog ka humanom. Sedam od 14 mutacija iz tabele 4.3.1.1 (R36K, F71S, T128S, T216I, S271P, E302K, M314L) povećavaju amplitude na F(0.295), a smanjuju amplitudu na F(0.055), sugerišući da su ove mutacije kritične za H1N1/N2 u povećanju afiniteta za humane interaktore. Važno je napomenuti da su tri mutacije

(R36K, T216I, S271P) od prethodno navedenih prisutne i kod svinjskih H1N1 virusa koji su inficirali ljude u SAD između 2005 i 2007 (slika 4.3.1.2b). ISM analiza je pokazala da bilo koja kombinacija sledećih mutacija F71S, T128S, E302K, M314L, koje su prisutne samo kod A/H1N1, smanjuje amplitude na F(0.055) i povećava na F(0.295). To ukazuje da ove četiri mutacije imaju važnu ulogu u efikasnoj infekciji ljudi virusom A/H1N1 i efikasnoj transmisiji ovog virusa s čoveka na čoveka. Sedam od 14 mutacija iz tabele 4.3.1.1 smanjuju amplitudu na F(0.055) i povećavaju amplitudu na F(0.295), 6 mutacija smanjuju amplitude na obe frekvence ili nemaju nikakav efekat, dok samo jedna (N168D) povećava amplitudu na F(0.055) i smanjuje amplitudu na F(0.295). Na osnovu izloženog bi se moglo zaključiti da su mutacije kod A/H1N1, koje predodređuje virus za interakciju sa humanim receptorom, više zastupljene nego mutacije koje predodređuju virus za interakciju sa svinjskim proteinom. Može se očekivati da će izolati A/H1N1 akumulirati dodatne polimorfizme u njihovim HA1 genima koji će dalje favorizovati interakcije sa humanim proteinom, što će prema ISM konceptu biti povezano sa povećanjem amplitude na frekvenci F(0.295) i smanjenjem amplitude na F(0.055). U cilju predikcije *hot-spot* mutacija koje bi favorizovale interakciju sa humanim receptorom, izvršeno je *in silico* skeniranje aminokiselinom alanin kompletnog HA1. Ova analiza je pokazala da bi mutacija rezidua 94D, 196D i 274D povećala amplitudu na kritičnoj frekvenci F(0.295) (slika 4.3.1.3a).

Pošto aminokiselina Asp ima najveću EIIP vrednost (tabela 3.6), supstitucija u bilo kojoj poziciji će povećati amplitudu na frekvenci F(0.295). Interesantno je, da je Asp (D) na pozicijama 94, 196 i 247 visoko konzerviran kod svih severno američkih svinjskih H1N1/N2 izolata i kod svih A/H1N1 HA1 gena. Jedini izuzetak su četiri humana izolata A/H1N1 iz Španije (A/Castilla-La Mancha/GP13/2009, A/Castilla-La Mancha/GP9/2009, A/Valencia/GP4/2009, A/Catalonia/P148/2009), dva izolata iz Italije (A/Italy/06/2009) i četiri izolata iz SAD (A/South Carolina/09/2009; A/South Dakota/05/2009; A/South Carolina/10/2009; A/Missouri/023/2009) koji imaju D274E mutaciju (slika 4.3.1.4). Ova mutacija značajno povećava amplitude na frekvenci F(0.295) i verovatno povećava afinitet za humane proteine.

a)

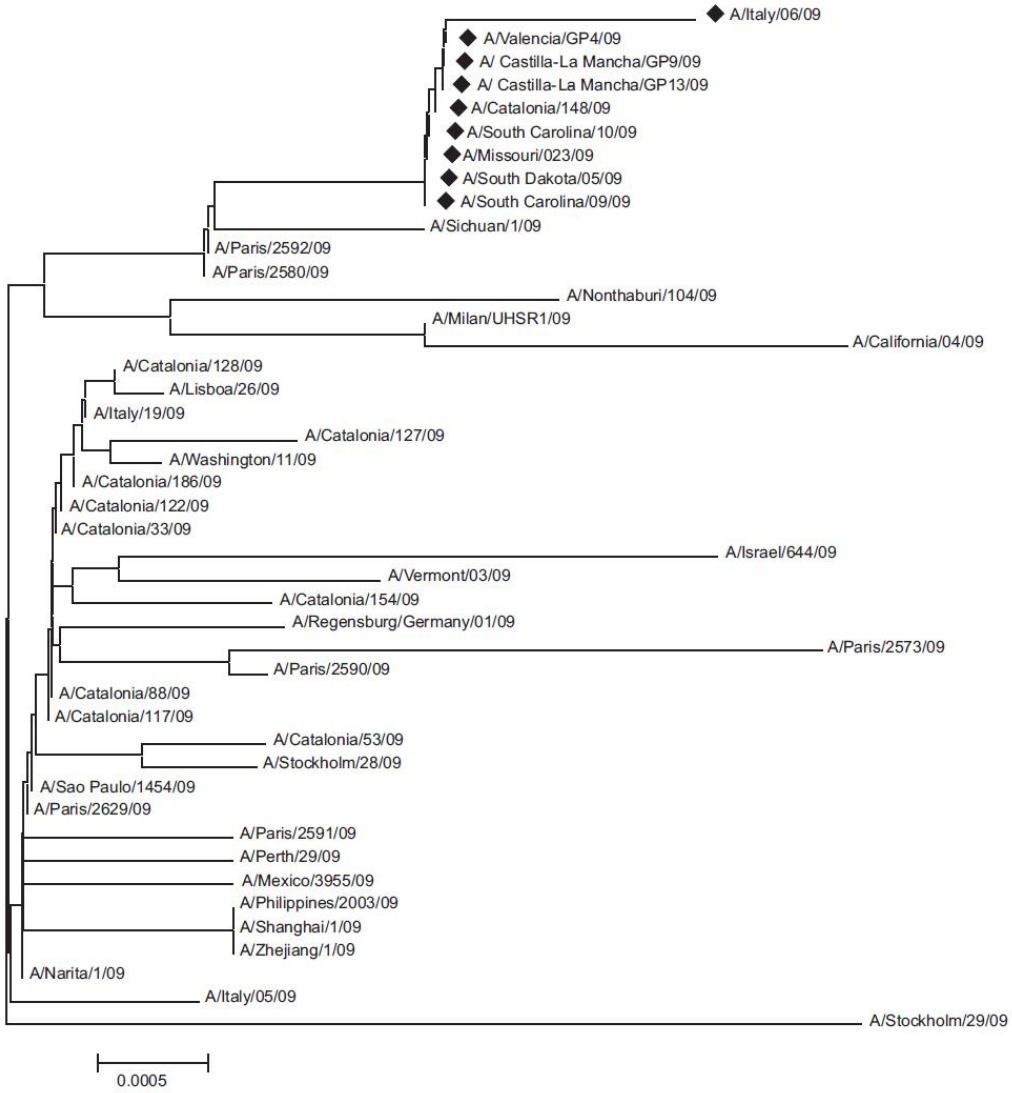


b)

```
Swine/Human     1 ----------------------------------NR-D----------------------
EPI177288       1 DTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDKHNGKLCKLRGVAPLHLGKCNIAGW
FJ985753        1 --------------------------------I-----------------------------

Swine/Human    61 L---------F-----------PN-----------N------H---------------A
EPI177288      61 ILGNPECESLSTASSWSYIVETSSSDNGTCYPGDFIDYEELREQLSSVSSFERFEIFPKT
FJ985753       61 ------------------------------------------------------------

Swine/Human   121 N------T-R-------Y—TN---R--------------IN-----N-E-----I----
EPI177288     121 SSWPNHDSNKGVTAACPHAGAKSFYKNLIWLVKKGNSYPKLSKSYINDKGKEVLVLWGIH
FJ985753      121 ------------------------------------------------------------

Swine/Human   181 --P--T---T----------------R—E-----T/A----N-A---------I-----T
EPI177288     181 HPSTSADQQSLYQNADAYVFVGSSRYSKKFKPEIAIRPKVRDQEGRMNYYWTLVEPGDKI
FJ985753      181 ------------------------------------------------------------

Swine/Human   241 ---------A------LK—S---------SI---D------N--------------V--
EPI177288     241 TFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPKGAINTSLPFQNIHPITI
FJ985753      241 --------------------------------E-------------------------

Swine/Human   301 -E-----------M------I------
EPI177288     301 GKCPKYVKSTKLRLATGLRNVPSIQSR
FJ985753      301 ---------------------------
```

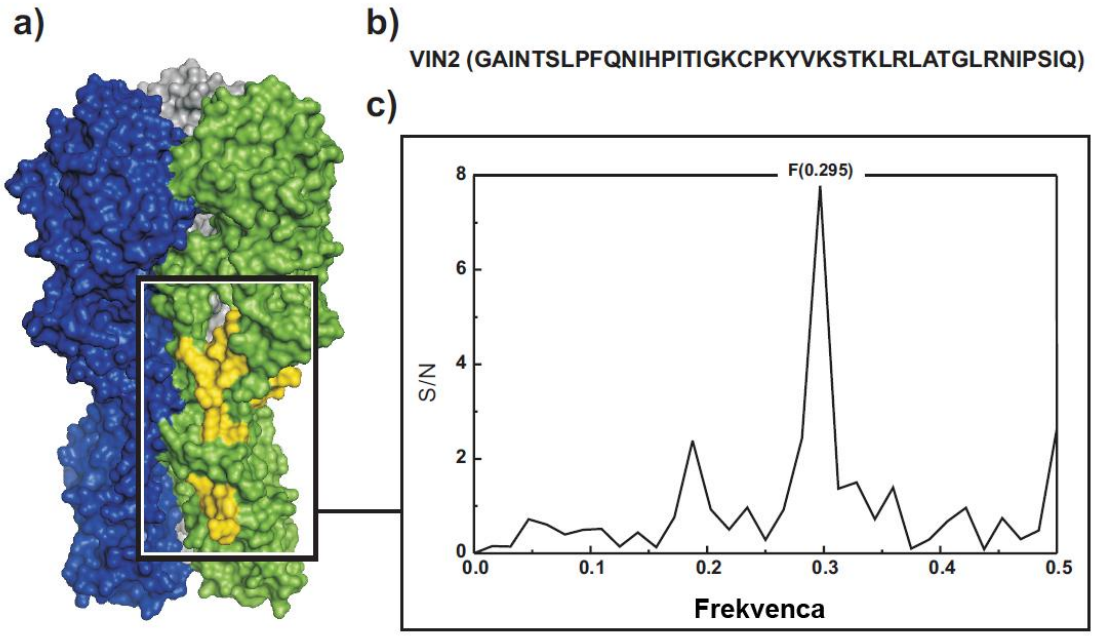
Slika 4.3.1.3. Identifikacija *hot-spot* mesta za mutacije koje mogu povećati A/H1N1/-receptor interakciju (a) Efekat supstitucije alaninom na amplitudu na frekvenci F(0.295): *In silico* alaninski sken HA1 sekvence ranih A/H1N1 izolata A/Mexico/4115/2009 (EPI177288). (b) Homologia između HA1 A/Castilla-La Mancha/GP13/2009 (FJ985753) (identičan izolatu iz SAD A/South Carolina/09/2009), svinjski H1N1 virus kojim su bili inficirani ljudi 2005-2007 i A/Mexico/4115/2009 (EPI177288).

Isti rezultat je dobijen skeniranjem sa bilo kojom drugom aminokiselinom sa izuzetkom Asp. Kao što se vidi na slici 4.3.1.3b, A/Castilla-La Mancha/GP13/2009 (FJ985753) koji je identičan sa A/South Carolina/09/2009 (GQ221794), se razlikuje od ranih meksičkih A/H1N1 izolata A/Mexico/4115/2009 (EPI177288) samo u pozicijama I32L i E257D. Aminokiseline L i I imaju istu vrednost EIIP (tabela 3.6), ali L>I supstitucija ne utiče na informacioni spektar. Suprotno ovome, EIIP vrednosti za aminokiseline D i E se značajno razlikuju, pa D>E mutacija povećava amplitudu na frekvenci F(0.295) za 15%, pri čemu ima mali efekat na strukturne osobine proteina pošto su obe aminokiseline negativno naelektrisane. Mutacija D274E nađena u SAD, Španiji i Italiji bi mogla odgovarati većoj adaptaciji A/H1N1 ka humanim receptorima.



Slika 4.3.1.4. Filogenija HA genskih sekvenci reprezentativnih izolata novog gripa H1N1. Izolati sa specifičnim D274E mutacijom su označeni crnim rombom.

Računarskim skeniranjem HA1 aminokiselinske sekvence A/H1N1 izolata je pokazano da glavni doprinos informaciji reprezentovanom frekvencom F(0.295) potiče iz domena koji je lociran na C-terminalnom kraju proteina koji obuhvata rezidue proteina 286 - 326 (označen VIN2). Slika 4.3.1.5 pokazuje IS i poziciju VIN2 domena u 3D strukturi A/H1N1 izolata A/California/04/2009. Treba naglasiti da je VIN2 konzerviran u svim A/H1N1 i da su dva od četiri polimorfizma (E302K i M314L), koji su identifikovani kao kritični za humanu infekciju, locirani unutar ovog domena. Značaj ovih polimorfizama je naglašen činjenicom da je domen 286–326 takođe visoko konzerviran kod HA1 svinjskih virusa (HA1 kod samo 30 of 500 svinja H1N1 iz UniProt baze ima mutacije u ovom domenu).



Slika 4.3.1.5. A/H1N1 HA trimer i detalji VIN2 regiona. IS i pozicije VIN2 domena (žuto) u 3D strukturi A/H1N1 izolata A/California/04/2009.

Relativna pozicija domena koji se vezuje za receptor i receptor targeting domena (VIN2) u 3D strukturi A/H1N1 HA1 je slična poziciji ova dva domena kod sezonskog gripa, a razlikuje se od pozicije ova dva domena kod H1N1 virusa iz 1918. godine [224]. Ovo sugeriše da je efikasnost interakcije između HA A/H1N1 i njegovog

receptora sličan sezonskom gripu H1N1, ali manje efikasan nego kod H1N1 virusa iz 1918.

Analiza HA1 proteina severno-američkih svinjskih izolata virusa H1N1 i H1N2 i novog A/H1N1, primenom EIIP/ISM platforme, ukazala je na razlike u parametrima koje određuju afinitet virusa za humani ili svinjski receptor. Utvrđene su aminokiselinske supstitucije F71S, T128S, E302K, M314L u HA1A/H1N1 koje su neophodne za interakciju sa humanim receptorom. Takođe je izvršena predikcija *hot-spot* mutacija u virusu A/H1N1 HA1 koje značajno favorizuju interakcije sa humanim proteinima. Jedna od ovih mutacija D274E je već pronađena u A/H1N1 humanim izolatima is Španije, Italije i SAD, ukazujući da se virus dalje adaptira na humanog domaćina. Takođe je pokazano da visoko konzervirani domen 286 - 326 u HA1 ima važnu ulogu u A/H1N1-interakciji sa receptorom i predstavlja potencijalni dijagnostički, terapeutski i vakcinski target.

4.3.1.2. Konzervirana svojstva hemaglutinina virusa H5N1 i humanih virusa influence: značaj u terapiji i kontroli infekcije

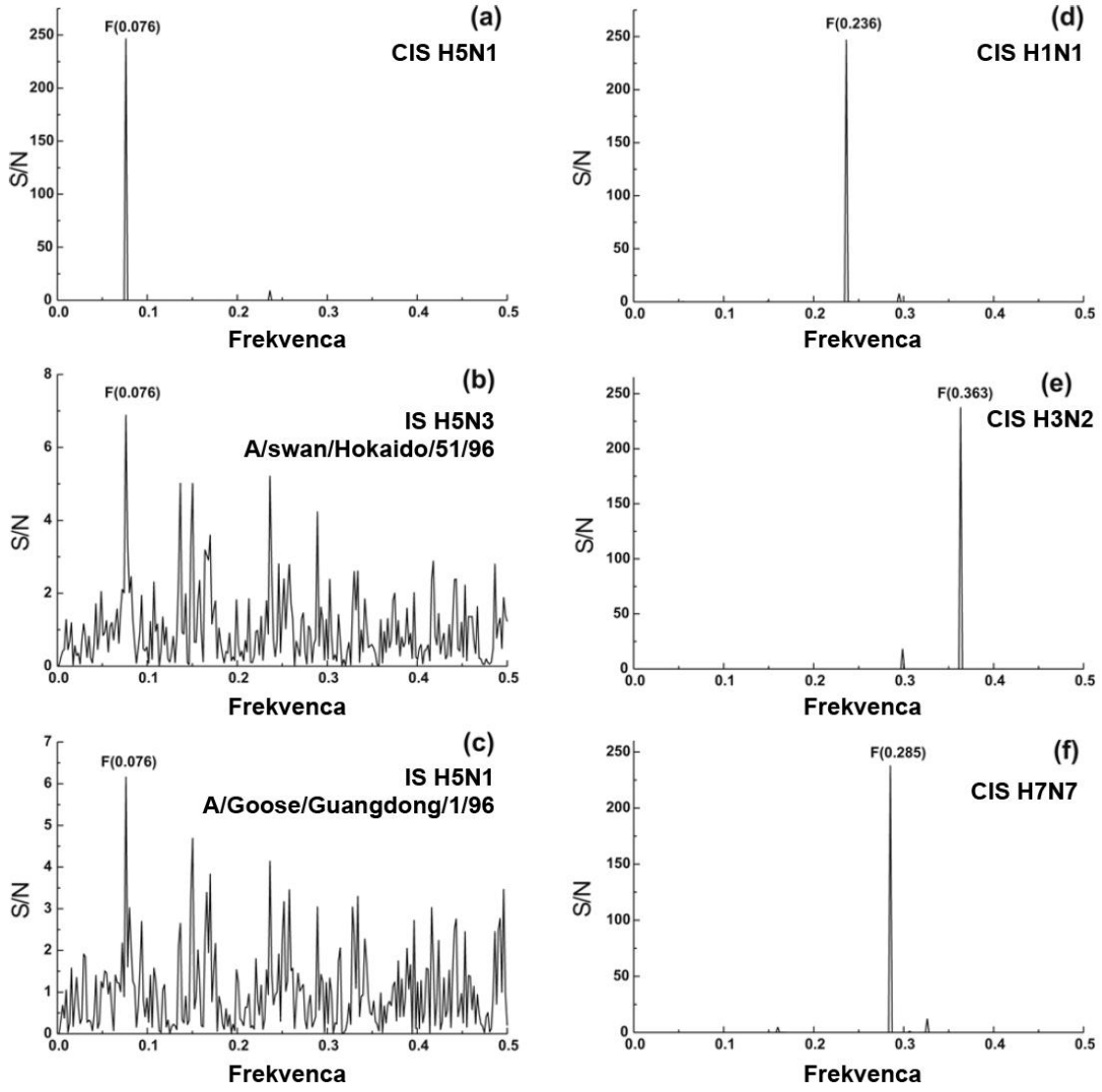
Epidemije izazvane visoko patogenim ptičijim influenca virusima (HPAIV) su stalna pretnja ljudskom zdravlju i svetskoj ekonomiji. Razvoj pristupa koji omogućavaju razumevanje značaja strukturnih promena izazvanim mutacijama, koje povećavaju mogućnost transmisije ovih virusa među ljudima, je od posebnog značaja. U radu su poređene informacione i strukturne osobine hemaglutinina (HA) značajne za interakciju virusa i receptora H5N1 virusa i drugih humanih subtipova virusa influence. Rezultati studije su omogućili bolje razumevanje virus/receptor interakcije i identifikaciju novih terapeutskih targeta.

Pošto promene povezane sa efikasnom transmisijom s čoveka na čoveka nisu razjašnjene, razvoj pristupa koji bi omogućio praćenje i razumevanje ovih promene je od velikog značaja za sprečavanja moguće pandemije.

Rezultati su objavljeni u sledećem radu [227]:
Veljkovic V, Veljkovic N, Muller CP, Müller S, Glisic S, ***Perovic V****, Köhler H. Characterization of conserved properties of hemagglutinin of H5N1 and human influenza viruses: possible consequences for therapy and infection control. BMC Struct Biol. 2009 Apr 7;9:21. doi: 10.1186/1472-6807-9-21.*

Da bi se identifikovala specifična informacija koja karakteriše dalekodosežnu komponentu protein-protein interakcija između H5N1 i njegovih potencijalnih receptora, primenom EIIP/ISM platforme, analiziran je HA1 protein H5N1 virusa influence. Kros-spektralnom analizom svih H5N1 HA1 aminokiselinskih sekvenci iz GenBank baze (1407 sekvenci) je utvrđeno da ovaj protein, iako visoko varijabilan, kodira konzerviranu informaciju reprezentovanu IS frekventnom komponentom F(0.076). Na slici 4.3.1.6a je prikazan konsenzus IS sa dominantnom frekvencom F(0.076). Prema konceptu ISM, ova informacija predstavlja dalekodosežnu komponentu protein-protein interakcije između HA1 i njegovog potencijalnog partnera - receptora. Slike 4.3.1.6b i c pokazuju IS HA1 H5N3 virus A/swan/Hokkaido/51/96, pretka HPA1 H5N1 subtipa i jednog od prvih H5N1 virusa izolovanih u Kini 2006. godine (A/Goose/Guangdong/1/96). Oba IS sadrže dominantan pik na istoj karakterističnoj frekvenci F(0.076) i HA oba analizirana virusa kodiraju istu informaciju kao H5N1 HA1 na slici 4.3.1.6a. Zatim su analizirane ISM metodom HA1 sekvence sezonskih H1N1 izolata (n = 29) i izolata H3N2 (n = 30), kao i izolata H7N7 (n = 30) iz različitih godina i iz različitih geografskih regiona. Konsenzus IS pokazuje karakterističan pik za svaku od ovih grupa F(0.236), F(0.363) i F(0.285) po prethodno navedenom redu, koji je različit od F(0.076) H5N1 HA.

Navedeni rezultati sugerišu da sekvence HA1 kodiraju specifične informacije za svaki podtip virusa influence.



Slika 4.3.1.6. ISM analiza HA1 proteina H5 influenca virusa. (a) Konsenzus IS HA1 svih H5N1 sekvenci iz GenBanka (n = 1407); (b) IS H5N3 (A/swan/Hokkaido/51/96), pretka H5N1, (c) prvi izolovani H5N1 virus (A/Goose/Guangdong/1/96); (d) konsenzus IS H1N1 (n = 30), (e) konsenzus IS H3N2 (n = 30) (f) konsenzus IS H7N7 (n = 30).

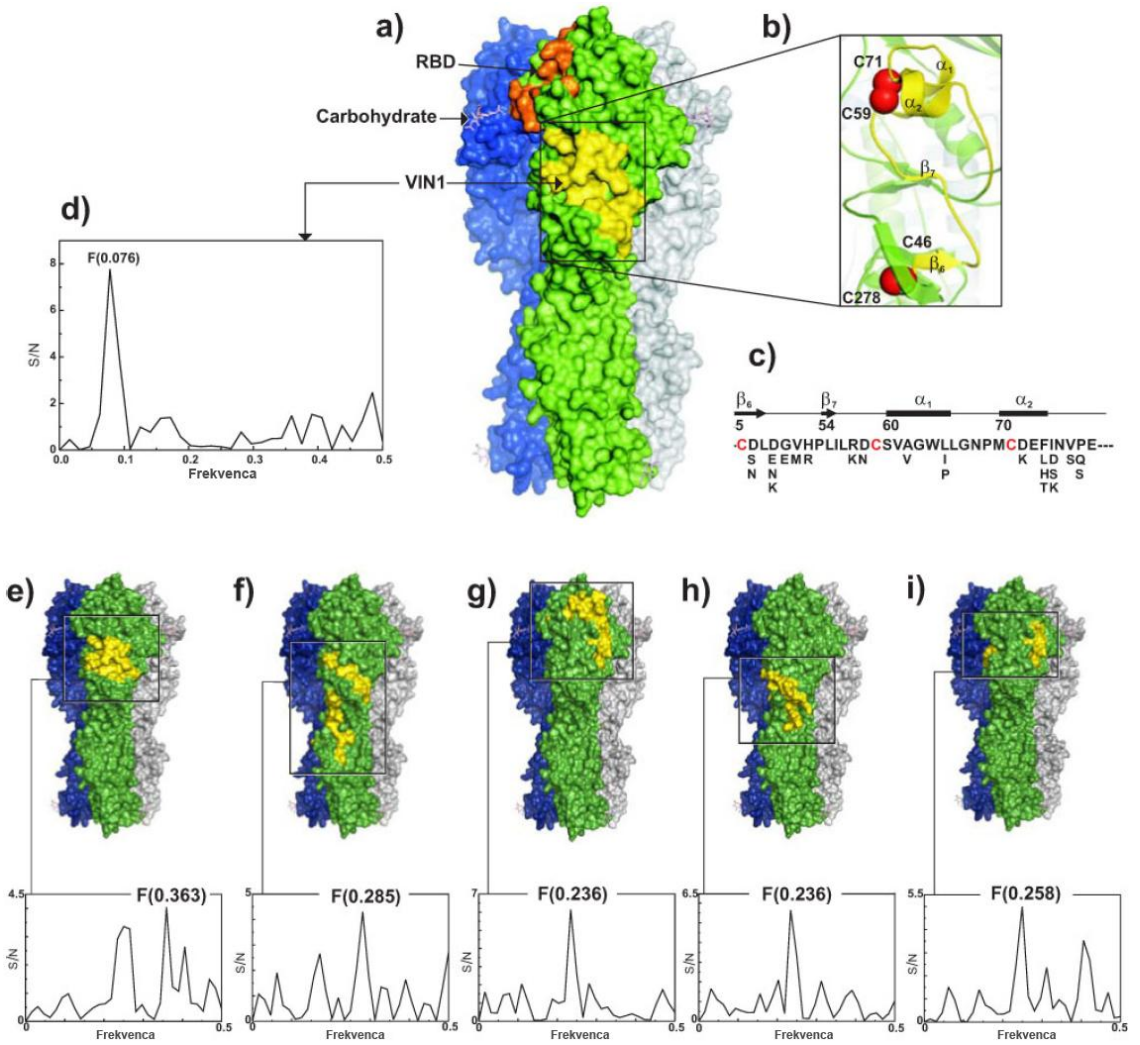
Računarskim skeniranjem primarne strukture H5N1 HA1 (primenom *SetSkener* programa EIIP/ISM platforme) identifikovan je domen HA1, označen kao VIN1, lociran na N terminalnom kraju proteina, koji obuhvata aminokiselinske rezidue 42–75 proteina HA (tabela 4.3.1.2, slika 4.3.1.7) i daje glavni doprinos informaciji reprezentovanoj frekvencom F(0.076). Ovaj domen H5N1 HA1 je visoko konzerviran kod svih H5N1 virusa. Peptid VIN1 je lociran unutar regiona E između aminokiselinskih pozicija 42 i 75, u jednom od pet glavnih antigenskih domena HA molekula. U 3D strukturi HA1

region E je lociran ispod globularne glave HA1 koja je uključena u vezivanju za receptor [228]. Ranije je pokazano da su domeni proteina, koji najviše doprinose određenoj IS frekvenci, direktno uključeni u protein-protein interakciju [172, 229]. Zato je pretpostavljeno da VIN1 domen ima važnu ulogu u prepoznavanju i targetingu između virusa i receptora. Stoga VIN1 može biti potencijalni target u terapiji infekcije virusom H5N1.

Većina mutacija koja određuje receptorski tropizam [230, 231] i izbegavanje imunskog odgovora je pronađena u globularnom delu HA1. Nasuprot tome, mutacije unutar regiona E su retke. Ovo ukazuje da varijabilni antigenski regioni A i B, locirani u globularnoj glavi HA1, mogu predstavljati imunski mamac koji štiti važan funkcionalni region E koji određuje konzervirane dalekodosežne osobine molekula. Slična strukturna organizacija je već publikovana kod HIV1 gp120 [232, 233], pri čemu je naglašeno da ona predstavlja veliku prepreku u razvoju AIDs vakcine [234-236].

Tabela 4.3.1.2. Domeni za prepoznavanje receptora HA proteina kod H5N1, H1N1, H3N2 i H7N7 influenca virusa.

Izolat	**Frekvenca**	**Pozicije**	**Sekvenca**
A/Hong Kong/213/03 (H5N1)	F(0.076)	42-75	CDLDGVKPLILRDCSVA GWLLGNPMCDEFINVPE
A/New Caledonia/20/99 (H1N1)	F(0.236)	262-295	SGIITSNAPMDECDAKC QTPQGAINSSLPFQNVH
A/New York/383/2004 (H3N2)	F(0.363)	57-90	QILDGENCTLIDALLGDP QCDGFQNKKWDLFVER
A/equine/Prague/56 (H7N7)	F(0.285)	28-61	GIEVVNATETVEQTNIPK ICSKGKQTVDLGQCGL
A/Egypt/0636-NAMRU3/2007 (H5N1)	F(0.236)	99-132	EELKHLLSRINHFEKIQII PKNSWSDHEASGVSS
A/South Carolina/1/18 (H1N1)	F(0.258)	87-120	NSENGTCYPGDFIDYEE LREQLSSVSSFEKFEIF



Slika 4.3.1.7. H5 HA trimer (PDB: 2ibx) i detalji VIN1 regiona. (a) Površina HA trimera, gde je svaki monomer različito obojen. Lokacija receptor vezujućeg domena (RBD) (narandžasto) i VIN1 (žuto) su naglašene samo za jedan monomer. (b) Trakasto predstavljanje VIN1 regiona (žuto). Atomi sumpora uključeni u stabilizaciju VIN1 regiona su označeni kao crvene sfere. Slike su napravljene u programu PyMol. (c) Sekundarna struktura i aminokiselinski prikaz H5 H1N1 regiona. Konsenzus sekvenca VIN1 regiona je pokazana zajedno sa mutacijama nađenim u 595 H5 HA sekvenci korišćenjem BioEdita. Cisteinski rezidui su označeni crveno. (d) IS VIN1 regiona. Domeni proteina H1N1, H3N2, H5N1, H7N1 i španske groznice identifikovani konsenzus IS (tabela 4.3.1.2) i njihove pozicije u 3D strukturi HA1 i IS peptidne sekvence: (e) A/New_York/383/2004 (H3N2); (f) A/equine/Prague/56 (H7N7); (g) A/Egypt/0636-NAMRU3/2007(H5N1); (h) A/New_Caledonia/20/99 (H1N1); (i) A/South_Carolina/1/18 (H1N1).

Na slici 4.3.1.7. su prikazani informacioni spektri peptida VIN1 i domeni koji su identifikovani pomoću konsenzus informacionih spektara H1N1, H3N2, H5N1 i H7N7 virusa (tabela 4.3.1.2), kao i pozicije ovih domena u molekulu HA.

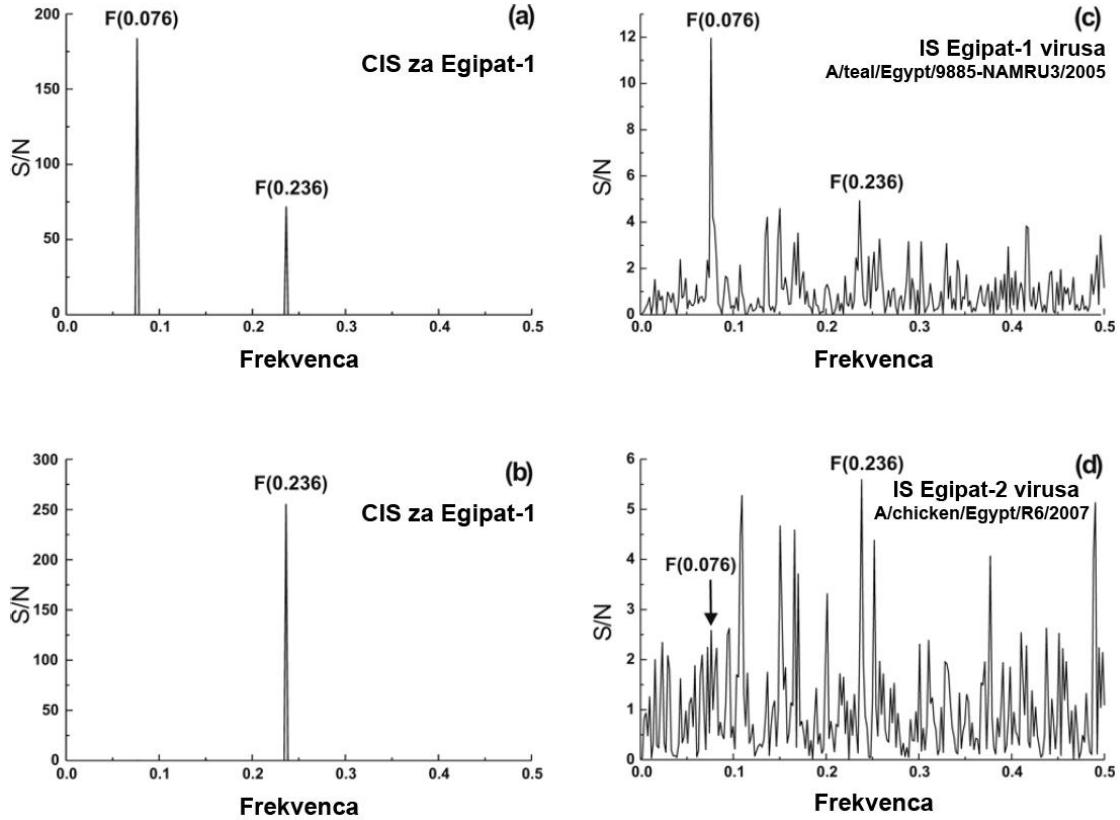
Iako je infektivnost virusa H5N1 za čoveka mala, u Egiptu je zabeležena pojava nekoliko klastera u kojima je transmisije virusa H5N1 bila sa čoveka na čoveka sa velikom stopom smrtnosti. H5N1 se efikasno replicira kod ljudi izazivajući stepen smrtnosti deset puta veći od pandemijskog gripa 1918. Stoga, ako bi infektivnost H5N1 postala slična sezonskom gripu, nastala bi pandemija katastrofalnih posledica. Najveća prepreka ovom jezivom scenariju je slaba transmisija H5N1 virusa s čoveka na čoveka, kojoj doprinosi malobrojnost 2,3 rezidua sijalinske kiseline na receptorima u epitelu humanog gornjeg respiratornog trakta i nesposobnost virusa da se efikasno replicira u ovom regionu.

ISM analiza 95 sekvenci HA1 H5N1 iz Egipta u periodu 2006-2007 pokazuje da je ove viruse moguće podeliti u dve grupe. Konsenzus IS prve grupe (Egipat 1), koja se sastojala od 55 izolata, ima dominantan pik na F(0.076) koji je karakterističan za HA1 H5N1 i drugi manji pik na F(0.236) karakterističan za H1N1 HA1 (slika 4.3.1.8a). Druga grupa Egipat 2 (slika 4.3.1.8b), koja obuhvata 40 H5N1 HA1, ima samo jedan dominantan pik na F(0.236) koji odgovara konsenzus IS H1N1 HA1 sa slike 4.3.1.6d. Slike 4.3.1.8c i d pokazuju reprezentativne IS pojedinačnih izolata iz obe grupe.

Od svih H5N1 virusa izolovanih u Egiptu tokom 2006. godine, 76% pripada grupi Egipat-1 a 24% grupi Egipat-2, dok 48% virusa izolovanih 2007. u Egiptu pripada grupi Egipat-1 a 52% grupi Egipat-2.

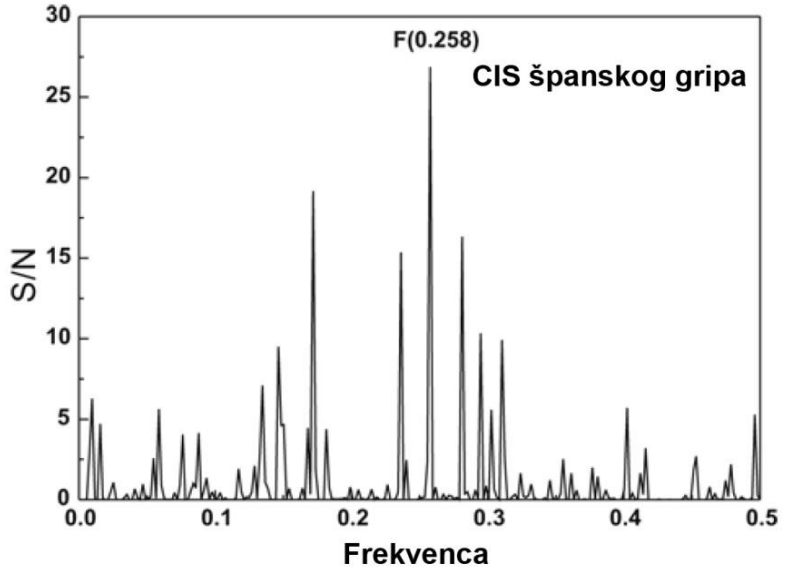
U ovom radu je pokazano kako ISM pristup identifikuje važne razlike između H5N1 virusa iz Egipta. Neki od njih imaju karakteristike kao i većina H5N1 izolata, dok jedna trećina ovih virusa ima karakteristike nađene kod humanih H5N1 sezonskih virusa, čija proporcija se povećala od 25% na 50% između 2006. i 2007. godine. Rezultati analize H5N1 izolata iz Egipta (slika 4.3.1.8) ukazuju na moguću evoluciju afiniteta virusa prema receptorima sličnim onima za H1N1 viruse, koji se efikasno repliciraju u gornjem disajnom traktu. Proteinski domen HA1 Egipta-2 koji je uključen u ovu pojavu odgovara aminokiselinskom domenu 99–132 (slika 4.3.1.7g). Uloga ovog domena u povećanoj infektivnosti virusa kod ljudi je još neobjašnjena. Interesantno je da su odgovarajući domeni odgovorni za targeting receptora kod virusa španskog gripa i

Egipta-2 mnogo bliže vezujućem mestu za receptor u HA1, nego kod svih drugih H1N1 i H5N1 virusa (slika 4.3.1.7e–i i tabela 4.3.1.2). Ova blizina bi mogla da ukaže na efikasniju interakciju virus/receptor kod ovih virusa influence.



Slika 4.3.1.8. ISM analiza HA1 proteina H5N1 virusa izolovanih u Egiptu 2006/2007. (a) konsenzus IS grupe Egipat-1 i (b) Egipat-2 (c) IS reprezentativnog Egipat-1 (A/teal/Egypt/9885-NAMRU3/2005) i (d) Egipat-2 virusa (A/chicken/Egypt/R6/2007).

Na kraju su poređene informacione karakteristike H1N1 pandemijskih izolata iz 1918. iz baze GenBank i sezonskih H1N1 izolata. Konzensus IS pandemijskih izolata iz 1918 karakteriše dominantan pik na frekvenci F(0.258) (slika 4.3.1.9), različit od F(0.236), karakterističnog za sezonske H1N1 izolate (slika 4.3.1.6d). Tabela 4.3.1.2 pokazuje domen HA koji odgovara frekvenci F(0.258). Kod modela A/South_Carolina/1/18 (slika 4.3.1.7i) pozicija domena koja odgovara karakterističnoj frekvenci se ne preklapa sa odgovarajućim domenom kod sezonskih H1N1 izolata, nego sa odgovarajućim domenom H5N1 virusa kod Egipta-2.



Slika 4.3.1.9. Konsenzus IS HA1 tri španska H1N1 virusa iz 1918.

Nedavno publikovani eksperimentalni rezultati ukazuju na funkcionalnu i imunološku ulogu H5 HA domena koji obuhvata peptid VIN1.

Da bi identifikovali mutacije koje povećavaju prepoznavanje H5 HA i SAα2,6Gal humanog tipa receptora, Su i saradnici su, poredeći HA A/chicken/Ffujian/1042/2005 sa izolatima identifikovanim i kod živine i kod ljudi u Kini, Hong Kongu, Tajlandu i Vijetnamu tokom epidemije u periodu 1996–2005 [237], otkrili šest tipova aminokiselinskih supstitucija. (K35R, D45N, D94N, K35R/D45N, K35R/45N/D94N, A247T) izvan receptor-vezujućeg domena HA, koji mogu povećati interakciju između H5 HA i humanog tipa SAα2,6Gal receptora. Može se videti da se kod 3 tipa ovih supstitucija D45E javlja samostalno ili zajedno sa jednom ili dve mutacije. D45E je locirana unutar peptida VIN1, a druge dve mutacije K35R i D94N su u blizini. Ovaj rad predstavlja prvi rad u kojem je objavljeno da prirodno nastale mutacije u regionu H5 HA koji uključuje peptid VIN1 imaju važnu ulogu u transmisiji virusa sa ptica na ljude. Važno je istaći da izolati iz Egipta sadrže sve ove mutacije osim K35.

Rezultati pokazuju da: (i) H5N1 HA1 kodiraju specifičnu informaciju predstavljenu IS frekvencom različitom od frekvenci za druge subtipove, (ii) ovoj karakterističnoj frekvenci najviše doprinosi visoko konzerviran N terminalni domen

HA1, (iii) drugi subtipovi influence kodiraju informacije koje odgovaraju drugim domenima HA: aminokiselinski rezidui 262–295 kod H1N1, 57–90 za H3N2, 28–61 kod H7N7 i 87–120 kod španske groznice, (iv) u Egiptu su se neki izolati virusa H5N1 adaptirali na receptor sličan receptoru za sezonski H1N1, što omogućava njihovu lakšu transmisiju s čoveka na čoveka.

4.3.2. Procena biološkog uticaja mutacija

Mutacije su promene u strukturi gena. Tačkaste mutacije su promene na nivou jednog nukleotida koje, kada nastanu u kodirajućem delu DNK, rezultuju nastanak kodona koji kodira drugačiju amino kiselinu. Ove promene na nivou DNK mogu dovesti do promene funkcije proteina i povezane su sa različitim bolestima.

Za procenu biološkog uticaja mutacija korišćeni su programi EIIP/ISM platforme: SeqEditor i FastaMutGen (Modul za obradu zapisa sekvenci); ProteinSpektar (Osnovni ISM modul); KrosSpektar (Modul za određivanje interaktora); AKSkener i SetSkener (Modul za određivanje interaktivnih domena); Mutacije i AKSkener (Modul za procenu biološkog efekta mutacija).

4.3.2.1. Procena biološkog efekta mutacija u molekulu lipoproteinske lipaze (LPL) kao faktor rizika za nastanak kardiovaskularnih bolesti

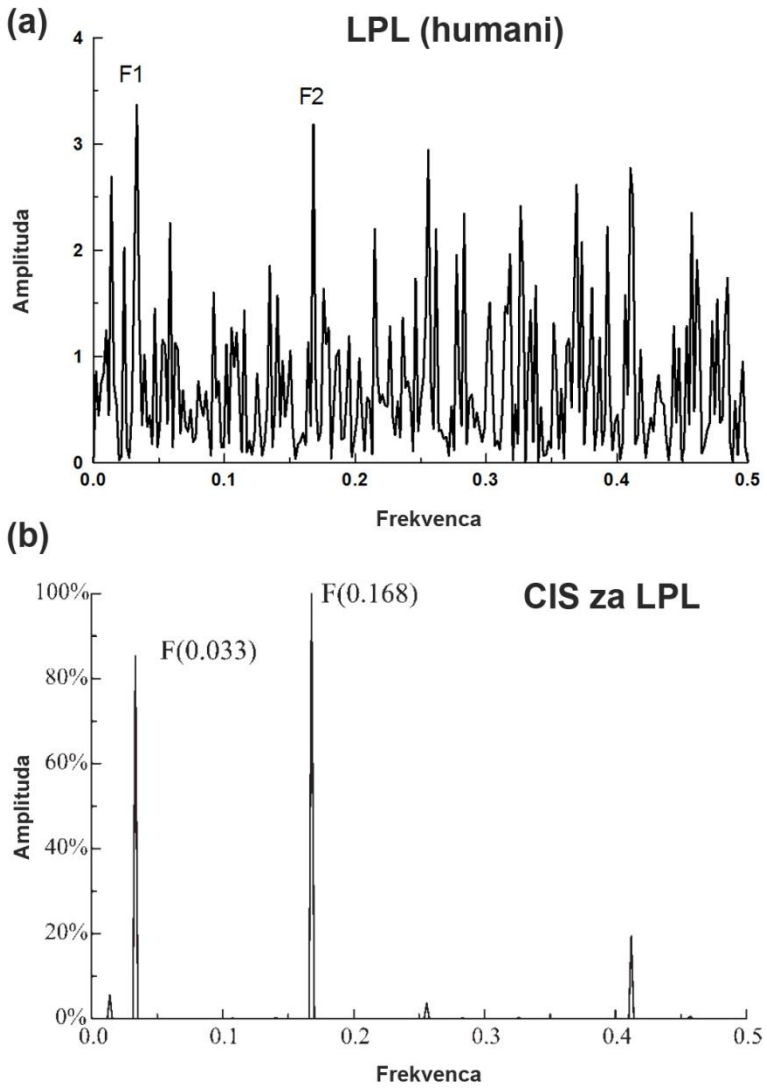
Lipoproteinska lipaza (LPL) ima centralnu ulogu u metabolizmu lipoproteina. LPL je odgovorna za hidrolizu triglicerida (TG) u lipoproteinima plazme i nastanak slobodnih masnih kiselina koje se mogu koristiti za izvor energije u mišićima ili se mogu deponovati u vidu masti u adipocitima (masnim ćelijama) [238]. U humanom genu za LPL je do sada dokumentovano više od 100 mutacija. Mutacije koje dovode do smanjene aktivnosti za LPL su povezane sa nepovoljnim lipidnim profilom i predstavljaju faktor rizika za nastanak kardiovaskularnih bolesti (KVB), koji je najčešći uzročnik morbiditeta i mortaliteta u razvijenim zemljama sveta.

Rezultati su objavljeni u sledećem radu [226]:
*Glisic S, Arrigo P, Alavantic D, **Perovic V**, Prljic J, Veljkovic N. Lipoprotein lipase: A bioinformatics criterion for assessment of mutations as a risk factor for cardiovascular disease. Proteins. 2008 Feb 15;70(3):855-62.*

Višestrukim poravnavanjem sekvenci (MSA) 14 LPL proteina iz kičmenjaka je ustanovljena homologija od 43.3%, koja je bliska homologiji slučajno izabranih sekvenci (37%). Kros-spektralnom analizom ovih sekvenci utvrđena su dva dominantna pika koji odgovaraju zajedničkim frekventnim komponentama F1(0.033) i F2(0.168) (slika 4.3.2.1b). Može se zaključiti iz ovih rezultata da primarne strukture molekula LPL kičmenjaka, uprkos ograničenoj homologiji, kodiraju zajedničke informacije koje su evolutivno konzervirane i odgovorne za biološku funkciju molekula LPL. Prema ISM konceptu, mutirane molekule LPL sa smanjenom biološkom aktivnošću karakteriše niža amplituda na karakterističnim frekvencama F1 i F2 u poređenju sa LPL molekulom divljeg tipa (eng. *wild type, WT*). Uzimajući ovu činjenicu u obzir, vrednosti amplituda koje odgovaraju frekvencama F1 i F2 u informacionom spektru humanog molekula LPL bile su osnovni parametar za praćenje uticaja mutacija na biološku aktivnost ovog proteina.

U tabeli 4.3.2.1 su prikazane 93 dokumentovane mutacije u humanom LPL molekulu koje delimično smanjuju ili potpuno ukidaju aktivnost LPL. Vrednost amplitude na jednoj ili obe frekvence je snižena u 75 od 93 analizirana mutirana proteina, koje karakteriše smanjena biološka aktivnost (80.6%). Daljom analizom ustanovljeno je da 38 od 45 (84.4%) mutiranih proteina, sa potpuno ukinutom aktivnošću LPL molekula, imaju smanjenu vrednost amplitude na jednoj ili obe frekvence u odnosu na divlji tip.

Iz navedenih rezultata može se zaključiti da promena amplitude na frekvencama F1(0.033) i F2(0.168) predstavlja pouzdan bioinformatički kriterijum za procenu efekta mutacija. Prema ovom bioinformatičkom kriterijumu, sve pojedinačne mutacije i njihove kombinacije, koje smanjuju amplitudu na jednoj ili obe frekvence, tretiraće se kao štetne po funkciju LPL molekula.



Slika 4.3.2.1. Poređenje informacionih spektara molekula LPL iz organizama na različitom stupnju evolutivnog razvoja. (a) Informacioni spektar humane LPL, (b) krosspektar LPL svih organizama.

Predloženi ISM kriterijum omogućava brzu i pouzdanu procenu efekta detektovanih mutacija posle sekvenciranja gena za LPL. Zato kriterijum može biti koristan za ranu identifikaciju osoba sa povećanim genetičkim rizikom da obole od KVB i preduzimanje preventivnih mera i terapija u cilju smanjenja rizika za nastanak KVB.

Tabela 4.3.2.1. Promena amplituda na F1(0.033) i F2(0.168) u informacionom spektru LPL uzrokovane mutacijama. Oznake u poljima su:

* (mutacije koje potpuno ukidaju enzimsku aktivnost LPL)

- (smanjenje amplitude na karakterističnoj frekvenci)

+ (povećanje amplitude na karakterističnoj frekvenci)

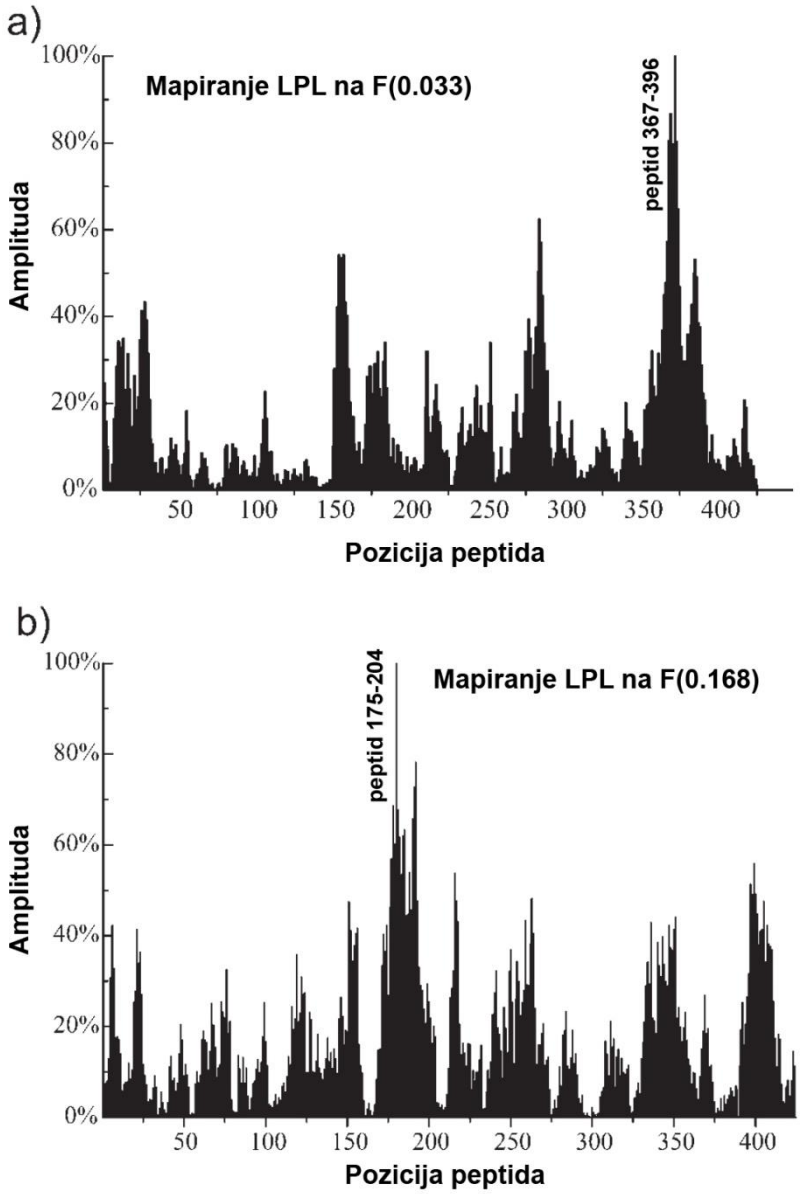
0 (bez promene amplitude na karakterističnoj frekvenci)

Mutacija	**F1**	**F2**	**Mutacija**	**F1**	**F2**	**Mutacija**	**F1**	**F2**
D9N	+	+	S172C	0	0	S259R*	-	-
D21V	-	+	D174V	-	-	A261T	-	-
C27X	-	-	A176T*	+	-	Y262H*	-	-
N43S*	-	+	D180E*	-	-	Y262X	-	-
H44Y	-	+	V181I	-	-	C264X	-	-
Y61X*	-	-	H183D*	+	-	S266P	-	+
W64X	-	-	H183N*	-	+	F270L*	-	-
V69L	+	+	H183Q	+	-	C283Y	+	+
A71T	-	-	G188E*	-	-	L286P	+	-
Y73X	-	-	G188R	-	-	Y288X*	-	-
R75S	+	+	R192Q	+	-	N291S	-	-
W86G	-	-	S193R	-	+	S298R	+	+
W86R*	+	+	G195E*	-	-	M301R*	+	-
A98T	-	+	D204E*	-	-	M301T	+	-
T101A	+	+	I205S*	+	+	Y302X*	-	-
G10R*	-	-	P207L*	-	+	L303P*	+	-
Q106X*	-	-	C216S*	0	0	L303F*	+	-
S132A*	+	-	I225T	-	-	S323C	0	0
H136R	-	-	C239X*	-	-	A334T*	+	+
G139S*	-	+	C239W*	-	-	S338F	-	-
G142E*	+	-	E242K	+	-	T352I	-	-
G154S*	+	-	R243C	-	+	L365V*	+	+

G154V*	+	-	R243H*	-	+	W382X*	-	-
D156G*	+	-	R243L	-	+	W382X	-	-
D156G*	+	-	S244T*	+	+	E410V	0	0
D156H*	-	-	I249T	-	-	E410K	-	-
D156N*	+	-	D250N*	+	-	C418Y	-	-
P157R*	-	+	S251C*	0	0	E421K	+	-
A158T	-	-	L252P*	-	+	C438S	0	0
E163D	-	+	L252V*	-	+	S447X	+	+
E163G	+	-	L252R*	-	+			
R170L	+	+	S259G	+	+			

Funkcionalno mapiranje LPL molekula

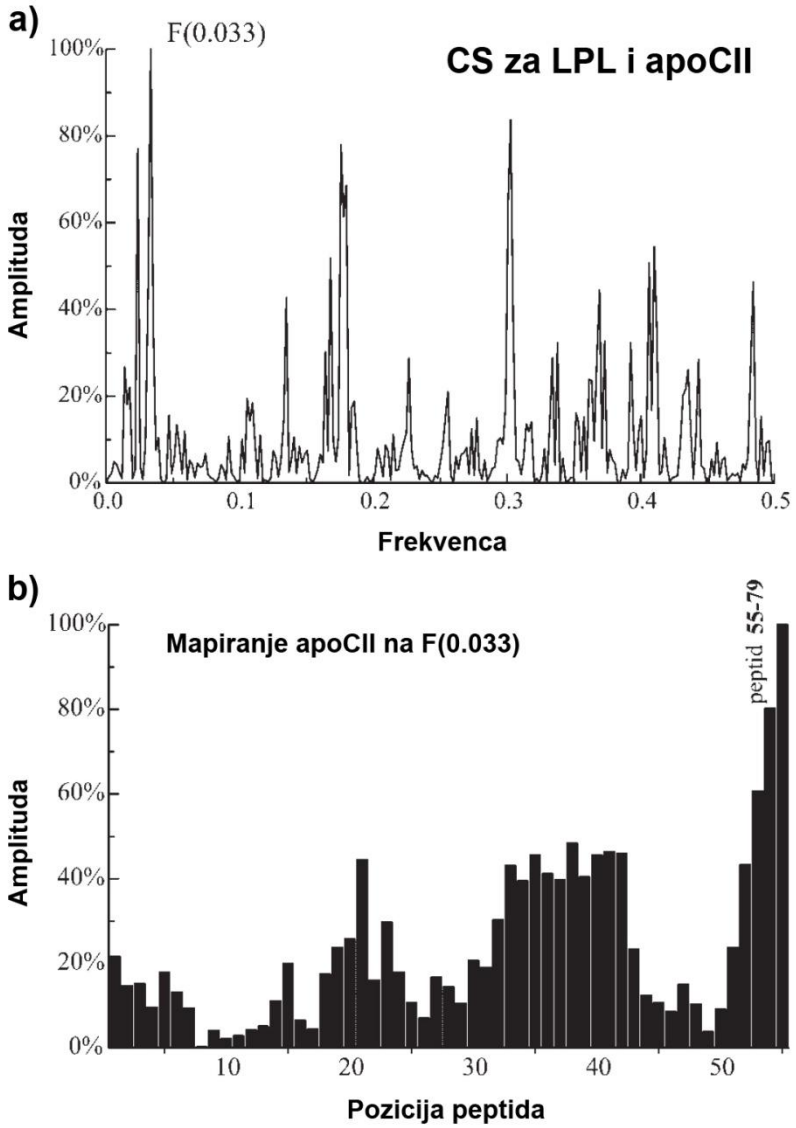
Da bi se odgovorilo na pitanje koje funkcije LPL su povezane sa frekventnim komponentama F(0.033) i F(0.168) u informacionom spektru ovog molekula, bilo je potrebno identifikovati prvo domene u primarnoj strukturi LPL, koji daju najveći doprinos informacijama reprezentovanim ovim frekventnim komponentama. Nakon identifikacije ovih domena, bilo je moguće na osnovu raspoloživih literaturnih podataka selektovati biološku funkciju LPL, koja se sa ovim domenima može povezati. Da bi se identifikovali domeni koji daju najveći doprinos frekventnim komponentama F(0.033) i F(0.168), primarna struktura molekula LPL je skenirana peptidima dužine 25 aminokiselina i za svaku poziciju peptida određena je vrednost amplitude na frekvencama F(0.033) i F(0.168). Rezultati ove analize, koji su prikazani na slici 4.3.2.2, pokazuju da najveći uticaj na frekventnu komponentu F(0.033) ima C-terminalni deo LPL molekula koji obuhvata aminokiseline 367-396. Prema literaturnim podacima ovaj region LPL je odgovoran za njegovu interakciju sa apoC2 [239]. Na osnovu ovoga bi se moglo zaključiti da je informacija kodirana u primarnoj strukturi LPL, koja odgovara frekventnoj komponenti F(0.033), odgovorna za interakciju LPL/apoC2. Da bi se potvrdio ovaj zaključak, na isti način je izvršena analizu molekula apoC2 kako bi se utvrdilo da li je ova frekventna komponenta takođe bitna za njegovu interakciju sa LPL. Na slici 4.3.2.3a je prikazan krosspektar za LPL i molekul apoC2.



Slika 4.3.2.2. Određivanje domena LPL molekula koji daju najveći doprinos informacijama reprezentovanim frekventnim komponentama F(0.033) i F(0.168).

Dominantan pik u ovom krosspektru odgovara frekvenci F(0.033), koji potvrđuje da je ova frekventna komponenta najbitnija za interakciju apoC2/LPL. Prema rezultatima prikazanim na slici 4.3.2.3b, C-terminalni deo apoC2 (aminokiseline 55-79) daje najveći doprinos ovoj frekvenci, što znači da bi trebao biti i najodgovorniji za njegovu interakciju sa LPL. Prema rezultatima Shena i saradnika [240], alfa heliks koji obuhvata aminokiseline 59–75 u primarnoj strukturi apoC2 predstavlja mesto njegovog direktnog vezivanja za LPL. Konačno, prema rezultatima Balasubramaniama i

saradnika [241], sintetički peptid, koji uključuje aminokiseline 56–79 apoC2, aktivira LPL u istoj meri kao i nativni molekul. Na osnovu svih ovih rezultata, može se tvrditi sa velikom izvesnosti da frekventna komponenta F(0.033) u informacionim spektrima LPL i apoC2 reprezentuje informaciju koja je odgovorna za interakciju ova dva molekula.



Slika 4.3.2.3. Određivanje domena apoC2 molekula koji je odgovoran za njegovu interakciju sa LPL. (a) krosspektar LPL i apoC2 molekula. (b) peptidi u primarnoj strukturi apoC2 koji daju najveći doprinos frekventnoj komponenti F(0.033) u informacionom spektru.

Prema rezultatima prikazanih na slici 4.3.2.2b, najveći doprinos informaciji koja odgovara frekvenci F(0.168), daje centralni deo molekula LPL. Ovaj deo obuhvata aminokiseline 170–216, pri čemu je dominantan peptid na poziciji 175–204. Prema eksperimentalnim rezultatima funkcionalnog mapiranja LPL [242], njegov centralni deo uključuje domen odgovoran za direktnu interakciju između monomera LPL i formiranje funkcionalnog dimera. Region označen kao „domen dimerizacije“ čija je funkcija eksperimentalno utvrđena [242] obuhvata aminokiseline 176–195 i gotovo se u celini preklapa sa peptidom 175–204 sa slike 4.3.2.2b, koji daje najveći doprinos formiranju frekventne komponente F(0.168) u informacionom spektru za LPL. Na osnovu ovih rezultata se može zaključiti da frekvenca F(0.168) reprezentuje informaciju koja je odgovorna za interakciju monomera LPL i formiranje funkcionalnog dimernog molekula.

U ovom radu je, primenom ISM analizom, izvršena strukturno/funkcionalna analiza LPL molekula, kojom je utvrđeno da primarna struktura LPL kodira evoluciono visoko konzerviranu informaciju odgovornu za dimerizaciju LPL i za interakciju sa kofaktorom apoCII.

4.3.2.2. *In silico* kriterijum za predviđanje efekata *missense* mutacija u p53 na negativnu povratnu spregu sa proteinom Mdm-2

Divlji tip tumor supresor proteina TP53 je ključna komponenta ćelijskih mehanizama za održavanje integriteta genoma i homeostaze tkiva, jer je njena biološka funkcija eliminacija i sprečavanje širenje abnormalnih ćelija. Kod ćelijskog stresa, posebno kod oštećenja DNK, TP53 zaustavlja ćelijski ciklus dok traje reparacija DNK [243], ili ukoliko je ispravljanje oštećenja nemoguće uvodi ćeliju u apoptozu [244]. Ove aktivnosti mogu biti poništene od strane mutacija u p53 genu, koje predstavljaju najčešći tip genetskih promena u mnogim ljudskim malignitetima, što ukazuje da gubitak aktivnosti TP53 igra važnu ulogu u humanoj kancerogenezi. Mdm2 protein predstavlja najznačajniji regulator TP53 aktivnosti koji je identifikovan do sada. Mdm2 formira autoregulatornu povratnu spregu u kojoj divlji tip TP53 aktivira transkripciju

gena mdm2, i povećava nivo proteina Mdm2 koji se vezuje za TP53 i pokreće brzu razgradnju TP53 [245].

Rezultati su objavljeni u sledećem radu [203]:
*Veljkovic N, **Perovic V**. In silico criterion for prediction of effects of p53 gene missense mutations on p53-Mdm2 feedback loop. Protein Pept Lett. 2006; 13 (8): 807-14.*

Sekvence TP53, Mdm2, MDMX, p53 citoplazmatični protein sličan Parkinu (Park) i glukokortikoidni receptor (GR) su preuzete iz SvissProt baze podataka sa identifikacionom brojevima P04637, Q00987, O15151, Q8IWT3 i P041501, respektivno. Za analizu mutiranih TP53 i mogućeg odnosa između njihovih informacionih karakteristika i fizioloških posledica, korišćeni su podaci iz baze podataka IARC (R9) TP53 mutacija koja sadrži sve p53 genske mutacije identifikovane u humanim kancerima, objavljene u literaturi od 1989 [202]. Sekvence TP53 polimorfizama su generisane promenom TP53 na aminokiselinskim pozicijama 47 i 72. Mutirane TP53 sekvence su generisane uvođenjem *missense* mutacija u TP53 sekvencijalno. Ove sekvence su podeljeni u tri skupa podataka navedenih u tabeli 4.3.2.2. Skup mutanata sa germinalnim *missense* mutacijama je uzet iz IARC podbaze podataka koja sadrži TP53 mutacije osoba sa porodičnom istorijom raka, dok su podaci o tačkastim i višestrukim *missense* mutacijama dobijeni iz baze IARC TP53 somatskih mutacija kod sporadičnih kancera.

Tabela 4.3.2.2. Broj sekvenci u bazi TP53 mutacija.

Vrsta podatka	**Broj sekvenci**
TP53 polimorfizmi	4
TP53 sa tačkastim *missense* mutacijama	12975
TP53 sa višestrukim *missense* mutacijama	707
TP53 sa germinalnim *missense* mutacijama	937

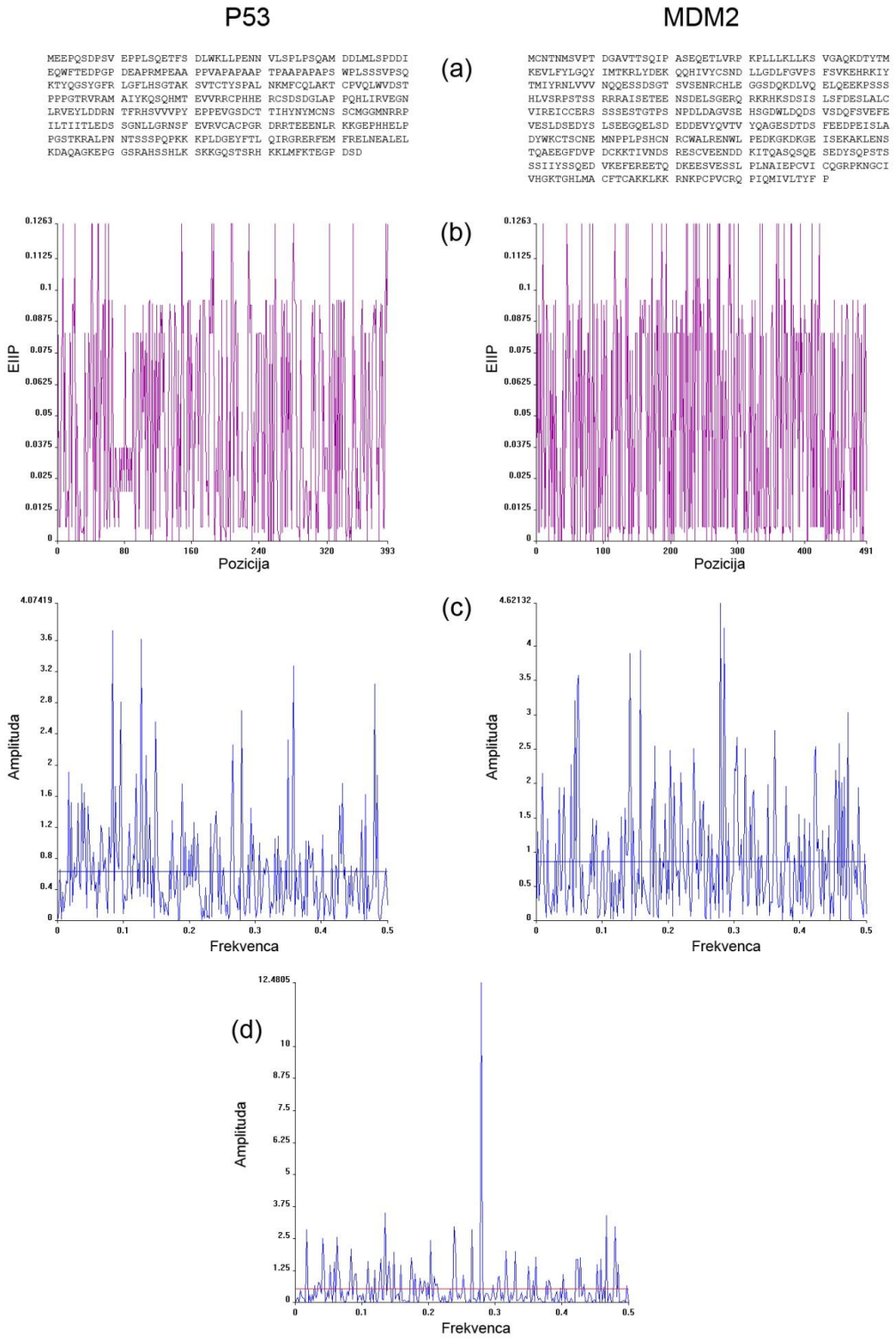
Za upoređivanje informacionih karakteristika mutiranih proteina i polimorfnih varijanti sa divljim tipom, primenjen je *Mutacije* program EIIP/ISM platforme. Ulazni podaci su bili TP53 sekvence, skup mutacija i karakteristična frekvenca spektra za

TP53-Mdm2 interakciju. Izlazni fajl je sadržao vrednosti amplituda za sve generisane TP53 sekvence koje odgovaraju izabranoj frekvenciji spektra. Prvo su analizirane TP53 polimorfne varijante, a zatim mutirane proteinske sekvence, kako bi se uporedile vrednosti amplituda sa odgovarajućim vrednostima funkcionalnih TP53 molekula.

Na slici 4.3.2.4 su prikazani informacioni spektri TP53 i Mdm2, kao i njihov kros-spektar, na osnovu kojih je identifikovana zajednička karakteristična frekvenca F(0.2793). Dakle, ova karakteristična frekventna komponenta u informacionom spektru TP53 odgovara dominantnoj informaciji koja je uključena u procesu prepoznavanja i targetinga Mdm2.

Dva polimorfizma pojedinačnih nukleotida (SNP), prolin u serin na poziciji 47, i arginin u prolin na poziciji 72, menjaju sekvencu TP53. Uprkos razlici u primarnoj strukturi, Pro72 i Arg72 se mogu smatrati konformacijski istim i divljim tipom [246], ali je potvrđeno da se TP53 varijante razlikuju u sposobnostima da indukuju apoptozu i suzbijanje rasta transformisanih ćelija [247]. Ovo je povezano sa diferencijalnom mitohondrijalnom lokalizacijom Arg72 varijante koja je povezana sa većim vezivanjem i degradacijom TP53 od strane Mdm2. Prema istraživanju Bougerada i saradnika kodon 72 polimorfizam predstavlja primer genetske predispozicije za rak, zbog svoje karakteristike da bude snažnije degradiran od Mdm2 nego divlji tip [248]. Iako polimorfizam u kodonu 47 nije bio predmet takve opsežne istrage do sada, eksperimentalni dokazi pokazuju da ovaj polimorfizam takođe utiče na sposobnost molekula da izazove apoptozu [249].

ISM analizom su izračunate amplitude A(0.2793) za različite varijante polimorfnih TP53. Minimalne i maksimalne vrednosti, date u tabeli 4.3.2.3, definišu interval u kojem se nalaze amplitude A(0.2793) funkcionalnih TP53 molekula. Dakle, interval (92.51% - 100.0%) je nazvan funkcionalni interval. Pretpostavljeno je da mutirani TP53 sa *missense* genetskom promenom koji smanjuju A(0.2793) ispod 92.51% granične vrednosti, imaju predispoziciju za povećanje degradacije u odnosu na divlji tip, a samim tim mutacije koja znatno snižavaju A(0.2793) mogu da utiču na disbalans regulacije povratne sprege, sa eventualnim uticajem na TP53 funkciju apoptoze.



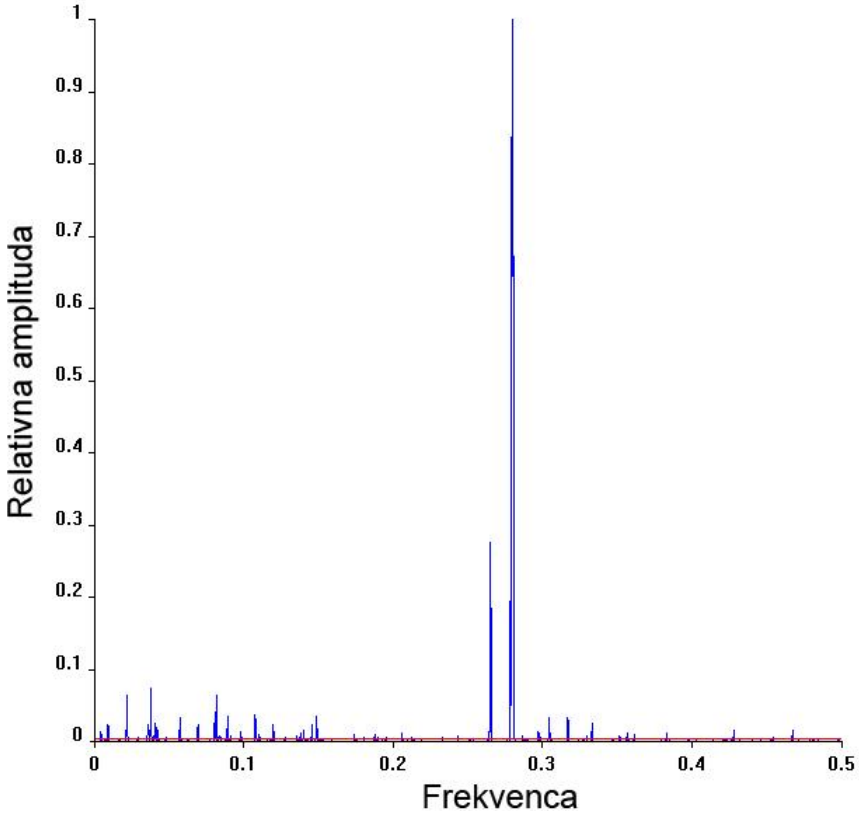
Slika 4.3.2.4. ISM analiza TP53 i Mdm2: (a) sekvence TP53 i MDM-2, (b) grafički prikaz odgovarajućih numeričkih sekvenci dobijenih zamenom svake aminokiseline sa odgovarajućom EIIP vrednosti (c) informacioni spektri za TP53 i Mdm2; (d) kros-spektralna funkcija spektara prikazanih u (c). Istaknuti pik označava zajedničku frekventnu komponentu.

Tabela 4.3.2.3. Relativne amplitude na karakterističnoj frekvenci 0.2793 u informacionim spektrima četiri polimorfnih varijanti humanog TP53 proteina.

TP53 varijante	**A(0.2793)**	**A(0.2793) [%]**
47S 72R	2.49845	92.51
47P 72R	2.57431	95.32
47S 72P	2.61518	96.84
47P 72P	2.70063	100.00

Da bi se istražilo da li neki drugi TP53-protein parovi imaju dominantnu frekvencu 0.2793 u svojoj kros-spektralnoj funkciji, analizirane su brojne sekvence TP53 interreagujućih proteina prijavljenih u literaturi. Izvođenjem krosspektralne analize, tri proteina sa posebnim spektralnim karakteristike su identifikovani: MDMX, glukokortikoidni receptor (GR) i citoplazmatični protein sličan parkinu povezan sa p53 (Park). Za sve njih postoje podaci da imaju direktnu i indirektnu ulogu u TP53-Mdm2 povratnoj sprezi u citoplazmi, i da kontrolišu TP53 apoptozu. Uloga MDMX, Mdm2 strukturno povezanog proteina, u regulaciji TP53 je veoma složena. MDMX interreagujući zajedno sa TP53 i Mdm2, predstavlja jedan od ključnih regulatora mreže [250]. Protein Park, delujući kao citoplazmatični sidro-protein u TP53 povezanim proteinskim kompleksima, je uključen u kontrolu p53 subćelijske lokalizacije, zatim i funkcije, a time i TP53 zavisne apoptoze [251]. Pored njih, GR formira kompleks sa TP53 i Mdm2 u citoplazmi i time izaziva povećanu degradaciju TP53 i GR u proteazomima [252].

U sledećem koraku je izračunat kros-spektar za sve navedene proteine. Zajednička frekventna komponenta 0.279 (slika 4.3.2.5) pokazuje sličnost njihovih spektara i može se smatrati kao konsenzus karakteristika ove funkcionalne grupe. Ovi rezultati snažno podržavaju pretpostavku o značaju ove frekventne komponente za Mdm2 zavisnu degradaciju TP53.



Slika 4.3.2.5. Kros-spektralna funkcija za TP53, Mdm2, MDMX, Park i GR. Jedini istaknuti vrh (pik) na frekvenci 0.279 u kros-spektralnoj funkciji označava samo jednu zajedničku frekventnu komponentu za sve analizirane proteine.

Specifične biološke i biohemijske osobine mutiranih TP53 proteina su predmet intenzivnog istraživanja, jer nisu sve tačkaste mutacije funkcionalno ekvivalentne, zbog čega nemaju svi mutirani proteini isti onkogeni potencijal [253]. Brojne studije ukazuju da TP53 status pacijenta može predvideti tok bolesti i odgovor na postoperativne terapijske intervencije, posebno na one terapije zasnovane na indukciji apoptoze u neoplastičnim ćelijama [254]. Uzeta zajedno, ova zapažanja ukazuju na potrebu za poboljšanjem karakterizacije TP53 mutiranih proteina. Cilj ovog istraživanja je bio da se izaberu mutanti koji su u stanju da poremete TP53 aktivnost u regulaciji povratne sprege.

Pokazano je u niz primera da su mutanti, koji su izgubili svoju primarnu funkciju, takođe imali smanjenu amplitudu na karakterističnoj frekvenci [173, 255, 174]. Utvrđene su vrednosti A(0.2793) za skupove *missense* mutacija u cilju

predviđanja onih mutacije koje će ozbiljno uticati na stopu TP53 degradacije od strane Mdm2.

Tabela 4.3.2.4 prikazuje skup registrovanih mutacija koje smanjuju vrednost amplitude spektra humanog TP53 na frekvenci 0.2793 ispod 92.51%. Već je pokazano od strane Venota i saradnika [256] da su TP53 mutirani proteini sa supstitucijom 281Gly bili više MDM2 degradirani od TP53 divljeg tipa, što je u skladu sa predviđanjima.

Tabela 4.3.2.4. Mutacije koje smanjuju vrednosti amplituda na frekvenci 0.2793 u spektru humanog TP53 ispod 92.51%.

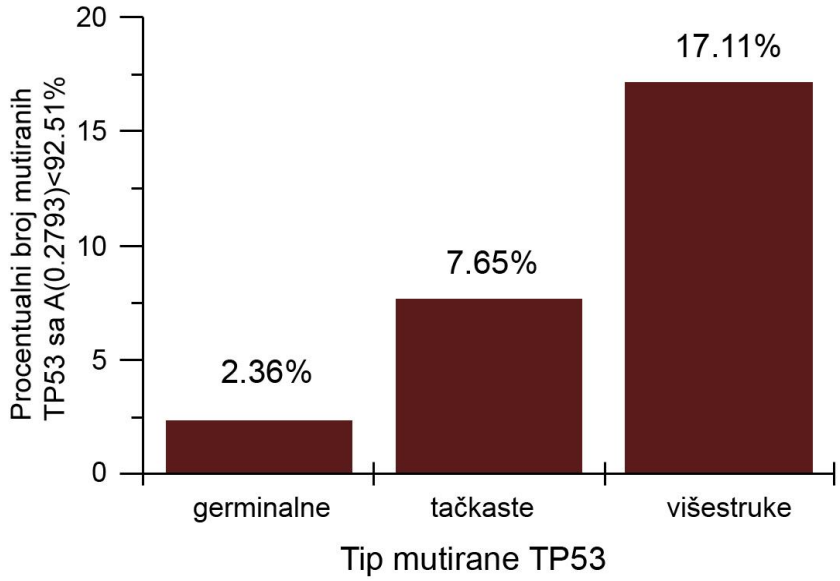
Pozicija	**Amino kiselina**	**Pozicija**	**Amino kiselina**	**Pozicija**	**Amino kiselina**	**Pozicija**	**Amino kiselina**
29	Asp	156	Gly	222	Arg	270	Leu
43	Ser	157	Asp	222	Thr	270	Val
48	Gly	157	Phe	225	Phe	272	Ser
68	Gln	161	Asp	231	Ile	272	Met
75	Arg	166	Leu	231	Asn	277	Gly
75	Ser	166	Gly	233	Asp	279	Arg
93	Met	168	Asp	236	Asp	281	Asn
102	Ile	168	Arg	238	Gly	281	Gly
107	Asp	170	Pro	247	Asp	281	Val
109	Leu	172	Asp	247	Thr	281	Glu
111	Arg	172	Phe	247	Ser	281	His
111	Gln	179	Asp	249	Ile	281	Ala
113	Leu	179	Arg	249	Asn	281	Tyr
113	Gly	181	Leu	249	Gly	281	Trp
113	Val	181	Gly	254	Asp	284	Ile
123	Ile	184	Asn	254	Phe	286	Asp
123	Asn	184	Val	254	Thr	286	Gln
129	Asp	184	His	254	Ser	293	Arg
134	Leu	184	Tyr	254	Met	297	Asp

134	Ile	193	Asp	256	Ile	306	Pro
134	Leu	193	Arg	256	Pro	308	Met
134	Val	197	Met	258	Asp	313	Asn
141	Gly	200	Ser	259	Asn	317	Leu
143	Met	202	Leu	259	Val	322	Arg
147	Asp	202	Gly	259	Glu	324	Glu
148	Asn	202	Pro	259	His	335	Gly
148	Val	202	His	265	Arg	338	Ile
148	Glu	209	Ile	265	Met	342	Leu
148	Ala	213	Leu	265	Gln	342	Pro
154	Asp	213	Gly	267	Gly	344	Arg
154	Ser	213	Pro	268	Ser	352	His
156	Leu	218	Met	270	Ile	365	Arg

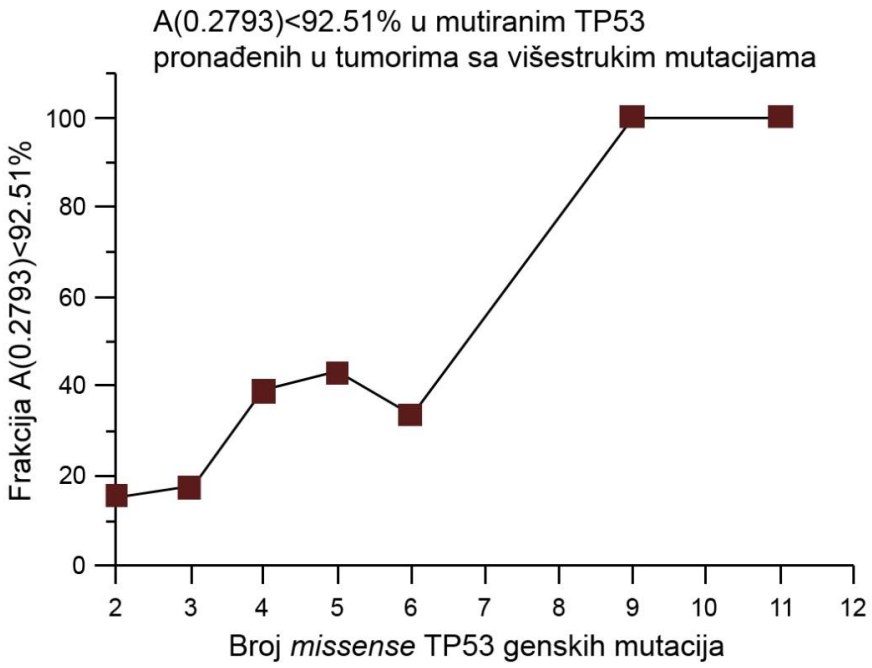
Sve je više prihvaćen koncept da je progresija ćelija sisara prema malignitetima evolutivni proces koji uključuje akumulaciju mutacija i na molekularnom i na hromozomskom nivou [257]. Dakle, TP53 mutirane sekvence sa višestrukim mutacijama predstavljaju grupu proteina sa povećanim onkogenim potencijal. Rezultati ISM analize predstavljeni na slici 4.3.2.6, otkrili su da je procenat sekvenci u grupi sa višestrukim mutacijama sa A(0.2793)<92.51% značajno veći od sekvenci sa tačkastim mutacijama. Ovaj rezultat ukazuje na značaj posmatrane osobine TP53 mutanata u procesu nastanka karcinoma.

Upoređeni su procentualni brojevi TP53 mutanata sa A(0.2793) ispod granične vrednosti 92.51% u zavisnosti od broja višestrukih *missense* mutacije gena p53 i pokazano je da povećanje broja mutacija povećava udeo mutanata sa kritičnom vrednosti amplitude. Ovi rezultati prezentovani na slici 4.3.2.7 eksplicitno ukazuju na sinergetski (ukupni) efekat p53 genskih višestrukih *missense* mutacija. Posmatrani sinergetski efekat može imati praktične implikacije na prognozu za osobe koje imaju mutacije. Naime, TP53 mutirani proteini čije su A(0.2973) vrednosti u funkcionalnom intervalu (92.51–100.0%), ali blizu donje granice 92.51%, imaju veće šanse da postanu inaktivni sa novo stečenim mutacijama. Nasuprot tome, molekuli čije su A(0.2973)

vrednosti u gornjem delu ovog intervala, su manje skloni da se inaktiviraju sa kasnijom *missense* mutacijom.



Slika 4.3.2.6. Distribucija TP53 sa relativnom vrednosti amplitude na karakterističnoj frekvenci 0.2793 ispod 92,51% u sledećim grupama: germinalne mutacije, somatske tačkaste mutacije, somatske višestruke mutacije.



Slika 4.3.2.7. Udeo TP53 proteina sa A(0.2793) manjom od 92.51% u grupi mutanata sa višestrukim mutacijama. Broj mutanata ispod praga se povećava sa brojem otkrivenih *missense* mutacija.

Višestruki genetski hit model raka [258] predviđa da bi normalni pojedinci trebalo da imaju stabilnu populaciju mutiranih ćelija koje su sklone raku, ali nekancerogene koje čekaju još genetskih mutacija. Jonanson i saradnici [259] su prijavili pet slučajeva p53 gena sa *missense* mutacijama u uzorcima kože zdravih osoba. Rezultati ISM analize pokazali su da su svi TP53 mutanti detektovani u normalnoj koži imali A(0.2793) u okviru funkcionalnog intervala (92.51 – 100.0%) (tabela 4.3.2.5). Ovaj rezultat je u skladu sa eksperimentalnim rezultatima [174] koji ne potvrđuju bilo kakvu sklonost za transformaciju ovih tačkastih mutacija u malim i odvojenim klonovima p53-mutiranih ćelija u normalnoj ljudskoj koži.

Tabela 4.3.2.5. Vrednosti A(0.2793) u informacionim spektrima mutiranih TP53, otkrivenih kod zdravih osoba, nalaze se u istom domenu četiri polimorfne varijante TP53.

Pozicija	**Mutacija**	**A(0.2793)**	**A(0.2793) [%]**
266	Gly → Glu	2.70213	100.05546
279	Gly → Glu	2.69800	99.90234
248	Arg → Gln	2.67406	99.01621
273	Arg → His	2.63264	97.48265
81	Thr → Ile	2.57057	95.18401

Kao posledica ovih analiza predlaže se sledeći bioinformatički kriterijum za procenu efekata TP53 mutacija na TP53-Mdm2 spregu: (i) mutirani proteini sa A(0.2973) ispod granične vrednosti 92.51% imaju predispoziciju za pojačanu degradaciju Mdm2 i (ii) mutirani proteini koji imaju vrednost A(0.2973) veću od granične vrednosti 92.51%, mogu se karakterisati kao normalni učesnici u povratnoj sprezi.

4.3.3. Filogenetska analiza

4.3.3.1. Primena novog filogenetskog algoritma za analizu proteina otkriva evoluciju H5N1 virusa influence ka efikasnoj humanoj transmisiji

Visoko patogeni virus ptičje influence tipa A podtipa H5N1 (eng. *Highly Pathogenic Avian Influenza Virus - HPAIV*) predstavlja veliku pretnju ljudskom zdravlju. H5N1 se prenosi sa ptica na čoveka sporadično i može biti fatalan za čoveka. Stopa smrtnosti HPAIV H5N1 kod čoveka iznosi preko 50% [260], što ga čini vrlo ozbiljnom pretnjom za ljudsku populaciju. Za sada H5N1 nema mogućnost efikasnog prenosa s čoveka na čoveka, ali tu prepreku može prevazići ukrštanjem sa H1N1 ili H3N2 virusima koji cirkulišu globalno. Među svim ljudskim H5N1 slučajevima objavljenim širom sveta u periodu između 2009 i 2011. godine, 58.6% se pojavilo u Egiptu [261], što ukazuje na široku mogućnost adaptiranja ovog virusa na čoveka u Egiptu.

Filogenetska analiza može otkriti sličnost šablona među virusima i utvrditi zajedničkog pretka virusa. Trenutni filogenetski pristupi za analizu evolucije virusa influence su zasnovani na višestrukom poravnavanju sekvenci. Ovi pristupi daju veoma korisne informacije o evoluciji virusa, ali imaju ozbiljne mane. Glavne slabosti su: (i) neosetljivost na poziciju mutacije i (ii) neuspešno analiziraju deleciju u sekvenci. Nedavni rezultati pokazuju da su pozicija i tip supstitucije, presudni za infekciju čoveka H5N1 virusom. Takođe je bitno navesti da pojedinačna delecija S129 u hemaglutininu H5N1 virusa, značajno povećava humani tropizam, odnosno povećava afinitet za humani receptor [262, 227].

Zbog navedenih prednosti ISTREE algoritma, opisanih u odeljku 4.1.2, i u cilju prevazilaženja nedostataka standardnih filogenetskih metoda, za filogenetsku analizu su korišćeni programi EIIP/ISM platforme: ProteinBazaSec i FastaMutGen (Modul za obradu zapisa sekvenci); ProteinSpektar (Osnovni ISM modul); KrosSpektar (Modul za određivanje interaktora); Mutacije (Modul za procenu biološkog efekta mutacija); ISMStablo i ISMGraf (Modul za filogenetsku analizu).

Rezultati su objavljeni u sledećim radovima [263, 264]:

Perovic VR. *Novel algorithm for phylogenetic analysis of proteins: application to analysis of the evolution of H5N1 influenza viruses. Journal of Mathematical Chemistry. 2013 Jun ;1-18.*

Perovic VR*, Muller CP, Niman HL, Veljkovic N, Dietrich U, Tosic DD, Glisic S, Veljkovic V. Novel Phylogenetic Algorithm to Monitor Human Tropism in Egyptian H5N1-HPAIV Reveals Evolution toward Efficient Human-to-Human Transmission. PloS one. 2013 Apr;8(4), e61572.*

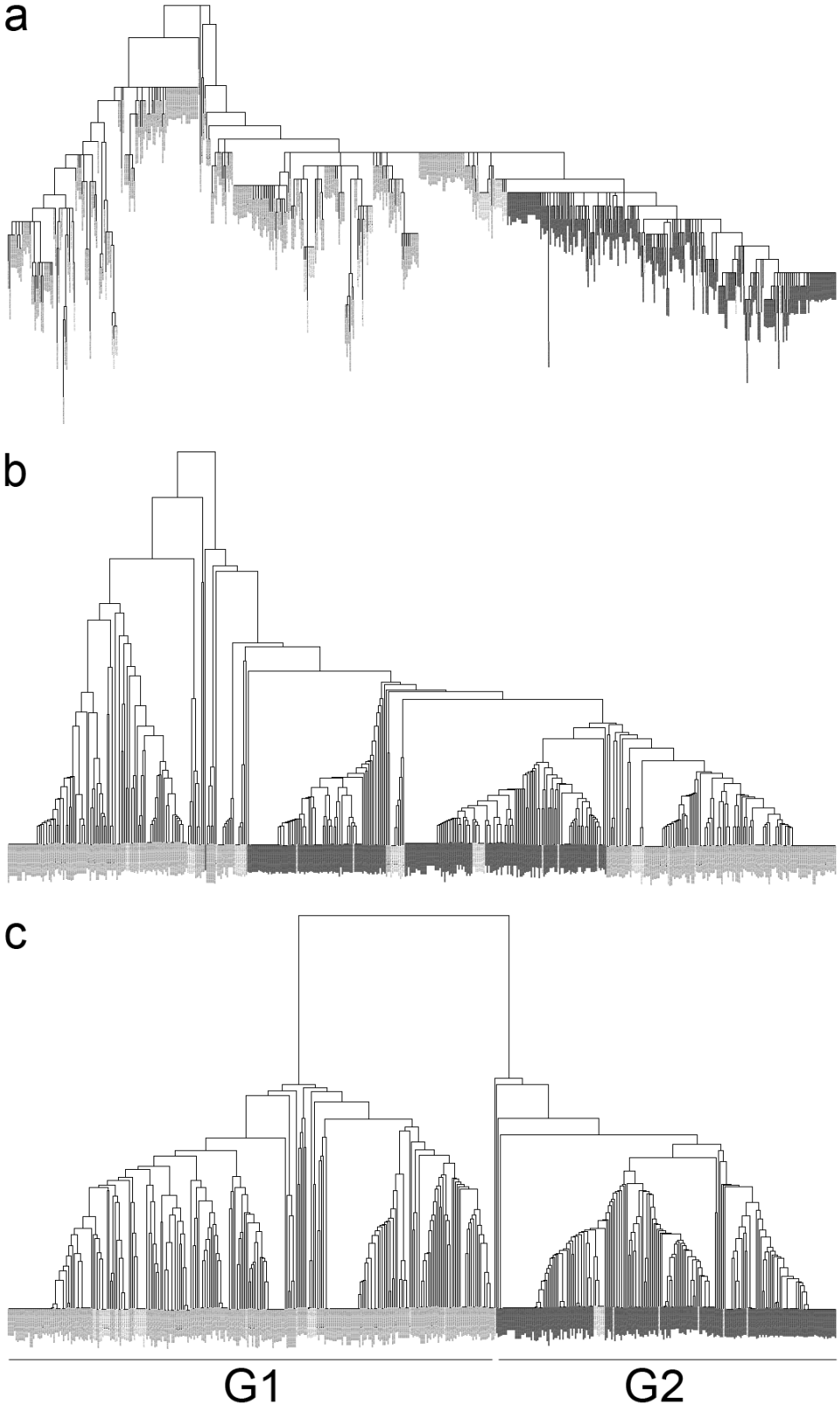
Svih 526 objavljenih sekvenci H5N1 influence A virusa izolovanog u Egiptu između 2006 i 2011. godine je preuzeto sa NCBI i GISAID baza [265], i analizirane su primenom standardnog filogenetskog i ISM zasnovanog filogenetskog pristupa.

Konvencionalna filogenetska stabla zasnovana na rastojanjima su izvedena korišćenjem softverskog paketa MEGA5 [197], dok je za ML stabla primenjen PHYML alat [46]. Za izračunavanje MSA sekvenci korišćeni su MUSCLE algoritam [198] i MEGA5 programski paket. ISM filogenetsko stablo je generisano primenom programa *ISMStablo* koji je deo modula za filogenetsku analizu EIIP/ISM platforme.

4.3.3.1.1. Analiza evolucije virusa H5N1 primenom ISM rastojanja na celom spektru

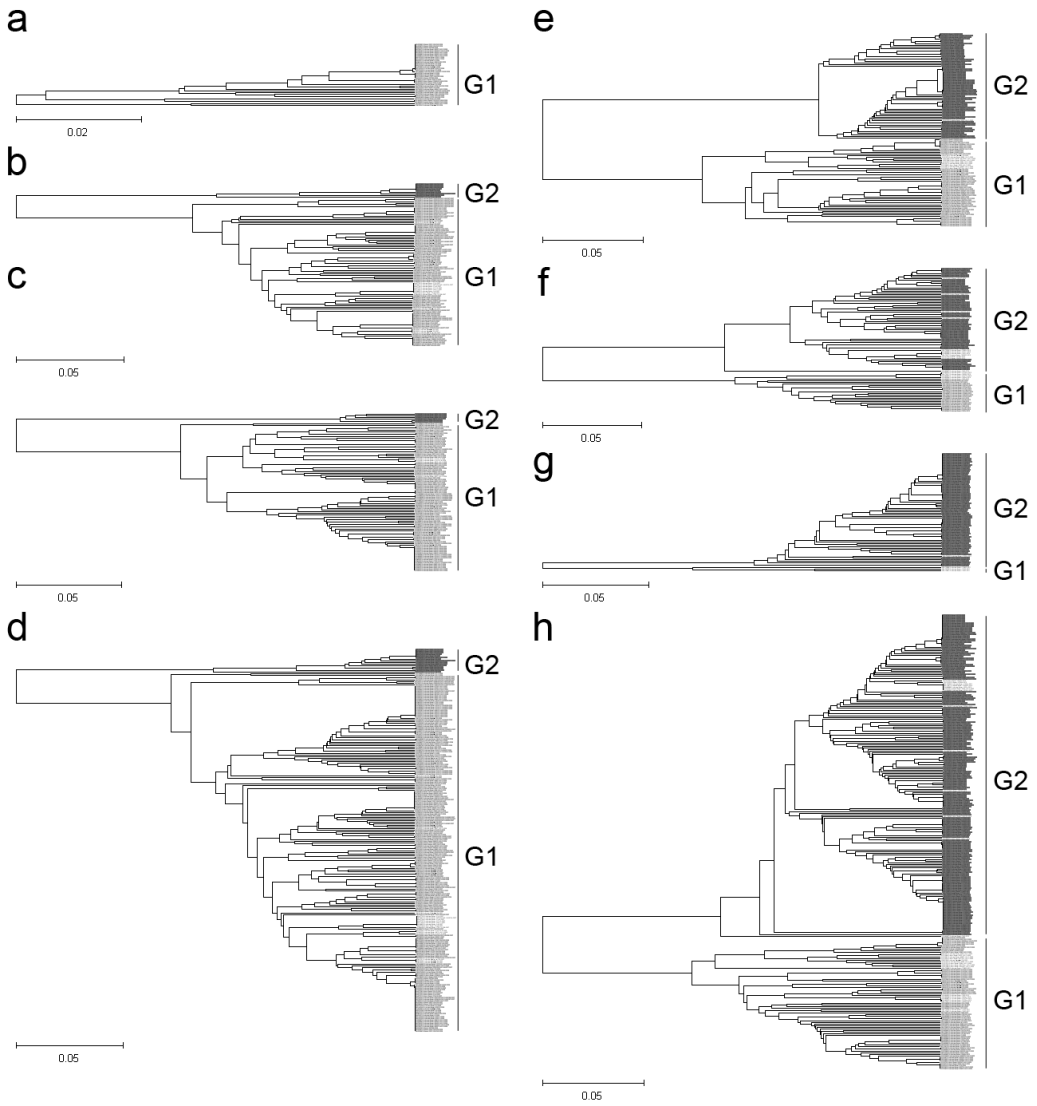
Za skup od 526 sekvenci H5N1 iz Egipta 2006-2011, generisana su filogenetska stabla, sa jedne strane standardnim metodama zasnovanim na višestrukom poravnavanju ML metodom i UPGMA metodom, a sa druge strane ISM zasnovanim pristupom, i upoređeni su rezultati (slika 4.3.3.1). Filogenetsko stablo zasnovano na ISM metodi pokazuje jasno grupisanje i odvajanje u dve odvojene grupe G1 i G2. Slična grupisanja postoje i u standardnim stablima, ali razdvajanja nisu toliko izražena kao u ISM stablu (slika 4.3.3.1).

Na slikama 4.3.3.1 i 4.3.3.2 su sekvence sa specifičnim aminokiselinskim ostacima D43, S120, (S,L)129, I151 obojene svetlo sivom, a sekvence sa specifičnim ostacima N43, (D,N)120, 129del, T151 tamno sivom bojom.



Slika 4.3.3.1. Filogenetska analiza H5N1 virusa izolovanog u Egiptu između 2006. i 2011. godine. Filogenetska stabla su konstruisana primenom: (a) ML metode; (b) UPGMA metode zasnovane na MSA; (c) ISM filogenetske metode.

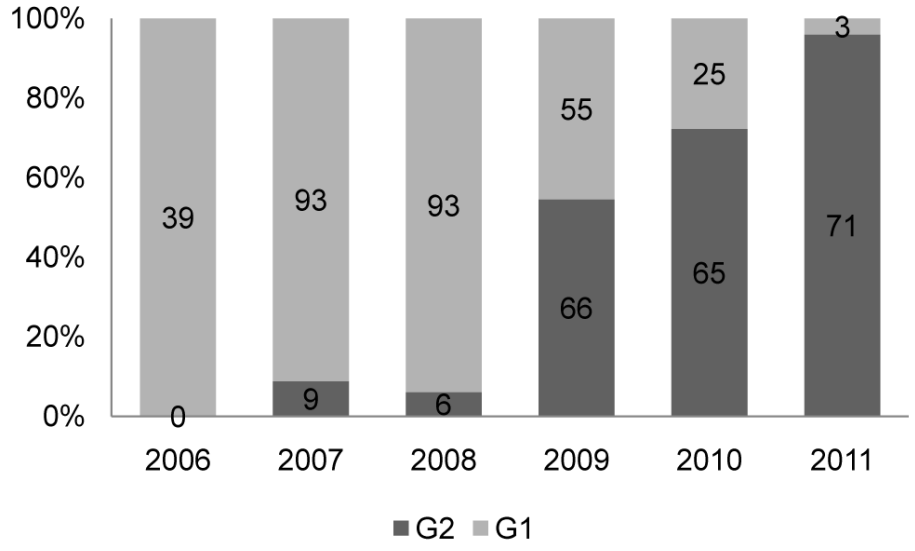
Detaljna ISM filogenetska analiza egipatskih H5N1 virusa za svaku godinu između 2006. i 2011. i za trogodišnje periode 2006-2008, 2009-2011 (slika 4.3.3.2), pokazuje da je broj sekvenci u G2 (tabela 4.3.3.1, slika 4.3.3.3) u konstantnom rastu od 0% u 2006 (slika 4.3.3.3a) do 95.95% u 2011 (slika 4.3.3.3g). Štaviše, analiza otkriva iznenadan skok u broju virusa u G2 posle 2008: od 6.06% u 2008 na 54.55% u 2009, i od 6.25% u periodu 2006-2008 na 41.25% u periodu 2009-2011 (slika 4.3.3.3d,h), što je u korelaciji sa iznenadnim povećanjem broja humanih slučajeva influence H5N1 u Egiptu posle 2008. godine, potvrđen od strane svetske zdravstvene organizacije (*World Health Organization WHO*) [266].



Slika 4.3.3.2. Detaljna ISM filogenetska analiza egipatskog H5N1 virusa za svaku godinu pojedinačno od 2006 do 2011. godine. ISM stablo za: (a) 2006; (b) 2007; (c) 2008; (d) period 2006-2008; (e) 2009; (f) 2010; (g) 2011; (h) period 2009-2011.

Tabela 4.3.3.1. Raspodela egipatskog H5N1 virusa između grupa G1 i G2, u ISM filogenetskom stablu konstruisanom za svaku pojedinačnu godinu od 2006 do 2011, trogodišnje periode 2006-2008 i 2009-2011, i za ceo period 2006-2011.

Godina	**G1**	**G2**
2006	39 (100 %)	0 (0 %)
2007	93 (91.18 %)	9 (8.82 %)
2008	93 (93.94 %)	6 (6.06 %)
2006-2008	225 (93.75 %)	15 (6.25 %)
2009	55 (45.45 %)	66 (54.55 %)
2010	25 (27.78 %)	65 (72.22 %)
2011	3 (4.05 %)	71 (95.95 %)
2009-2011	83 (29.12 %)	202 (70.88 %)
2006-2011	309 (58.75 %)	217 (41.25 %)



Slika 4.3.3.3. Raspodela egipatskog H5N1 virusa između grupa G1 i G2, u ISM filogenetskom stablu konstruisanom za svaku pojedinačnu godinu od 2006 do 2011 (tabela 4.3.3.1).

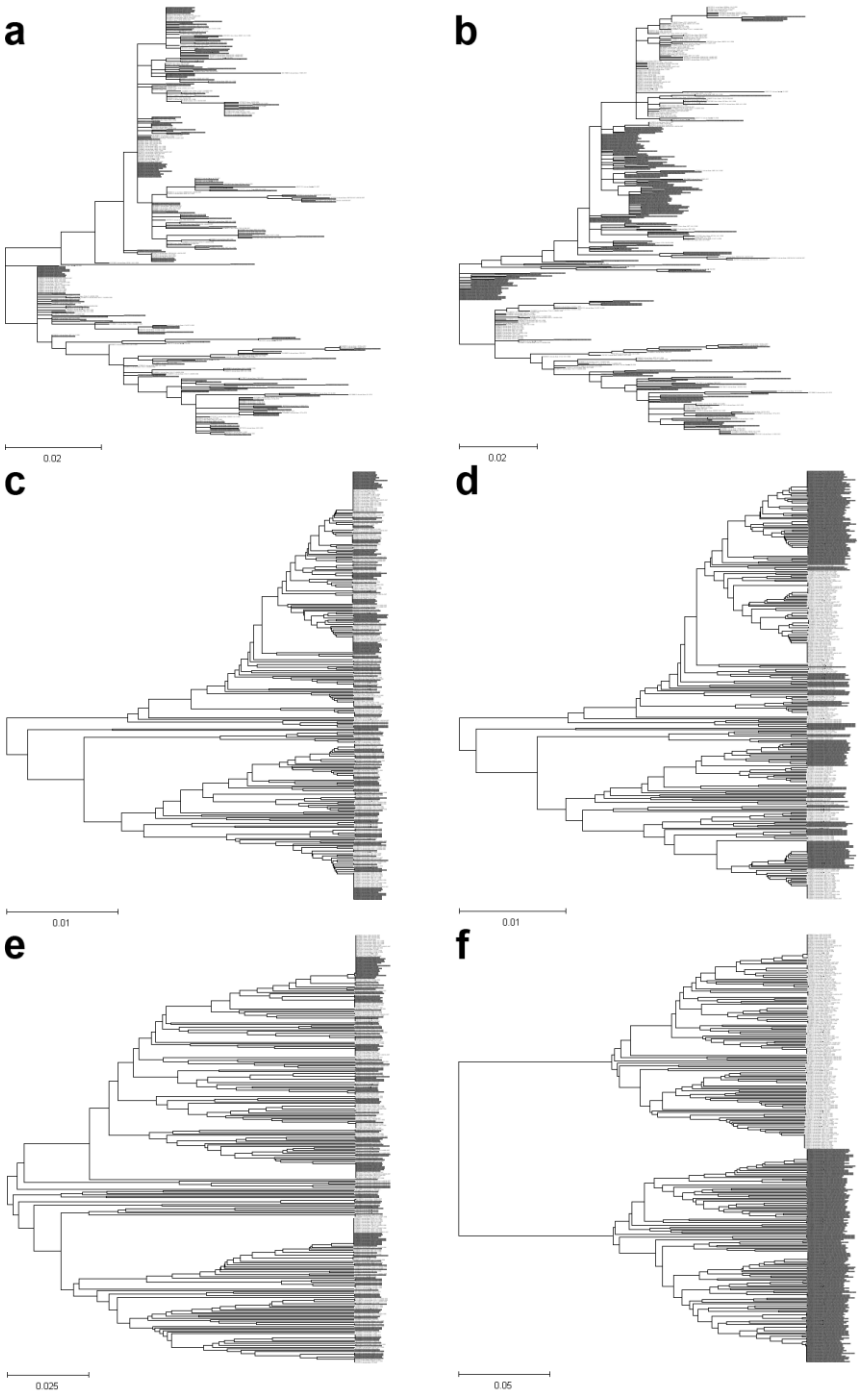
Analiza homologije sekvenci između grupa G1 i G2 otkrila je četiri specifične aminokiselinske pozicije (tabela 4.3.3.2): D43, S120, (S,L)129, I151 u G1, i N43, (D,N)120, 129del, T151 u G2. Povećanje humanih infekcija H5N1 virusom u Egiptu od 2009. godine, koje je u korelaciji sa naglim rastom G2 grupe u 2009. godini, sugeriše da su ove četiri mutacije bile bitne za povećanje humanog tropizma i pandemijskog potencijala ovih virusa.

Tabela 4.3.3.2. Aminokiselinski ostaci sa procentom njihove prisutnosti, koji su specifični za H5N1 HA1 u grupama G1 i G2.

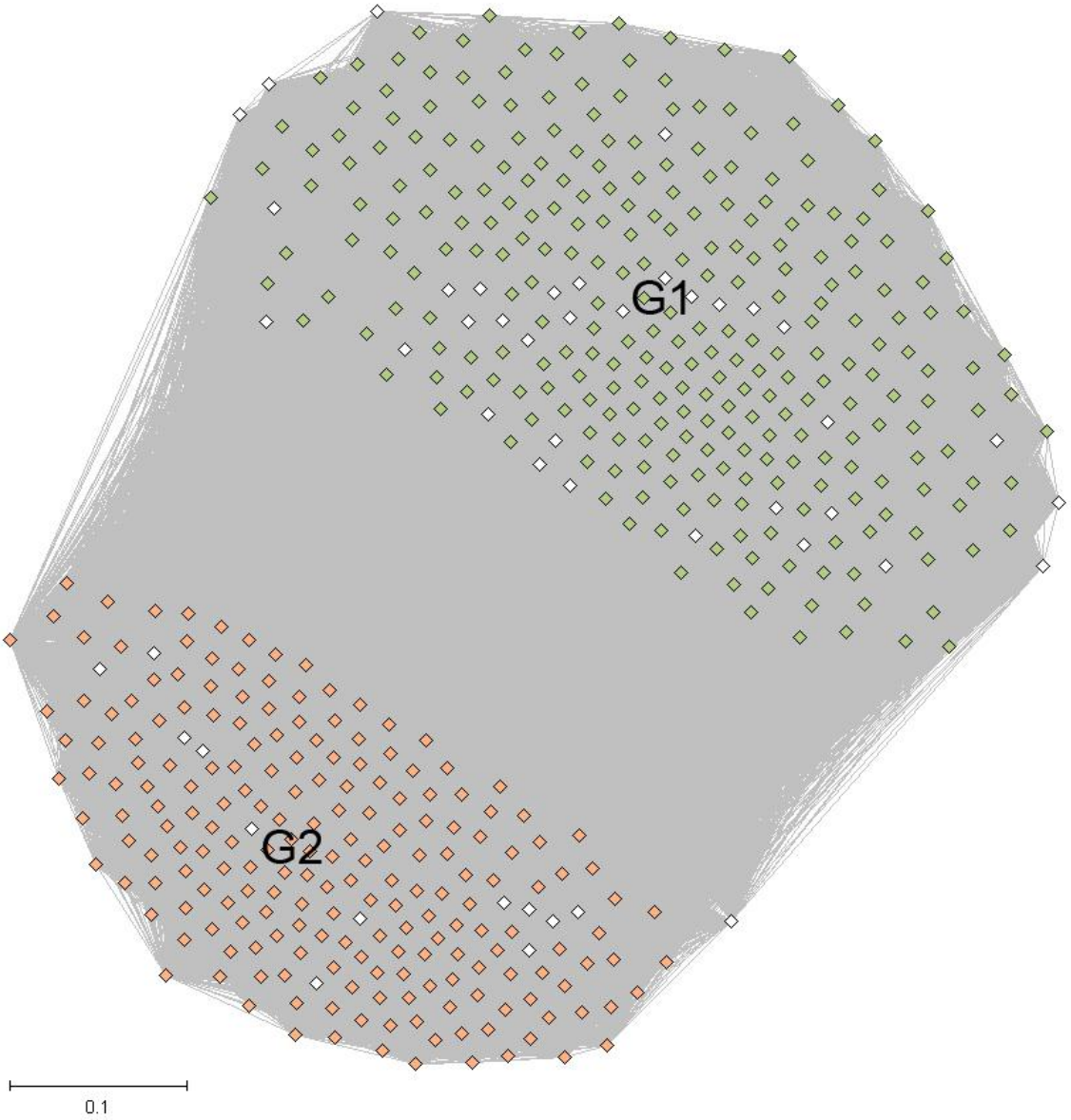
	Grupa G1	**Grupa G2**
Specifične aminokiseline	D43 (99%) S120 (94%) (S,L)129 (98%) I151 (92%)	N43 (98%) (D,N)120 (94%) 129del (99%) T151 (99%)

Za upoređivanje osetljivosti ISM filogenetske analize sa standardnim filogenetskim pristupom na mutacije i delecije, selektovano je svih 309 sekvenci iz G1 za test grupu. Pola od njih (svaka druga) je mutirano ubacivanjem karakterističnih mutacija iz G2 (43N, 120D, 129del, 151T). Za razliku od standardnog pristupa, u ISM filogenetskom stablu se vidi jasno grupisanje i odvajanje mutiranih od nemutiranih HA1 sekvenci (slika 4.3.3.4). Ovaj rezultat pokazuje prednost ISM filogenetskog pristupa nad standardnim filogenetskim metodama, u smislu osetljivosti na mutacije i delecije koje su esencijalne za biološku ulogu proteina.

Pored ISM filogenetskog stabla svih egipatskih H5N1 virusa (slika 4.3.3.1c), generisan je graf na osnovu modela elektro-opruga primenom programa *ISMGraf* , gde je svaki čvor jedna H5N1 sekvenca, grane su ISM rastojanja između sekvenci definisana na celom spektru kao kod generisanja ISM stabla, a za metodu konstrukcije je izabran efikasan algoritam simuliranog kaljenja (slika 4.3.3.5). Kao i kod ISM filogenetskog stabla, na grafu se vidi jasno odvajanje u dve grupe koje odgovaraju grupama G1 i G2 u ISM stablu, gde su čvorovi grafa, odnosno sekvence sa G1 karakterističnim aminokiselinama označene zelenom bojom, a sekvence sa G2 kiselinama označene narandžasto.



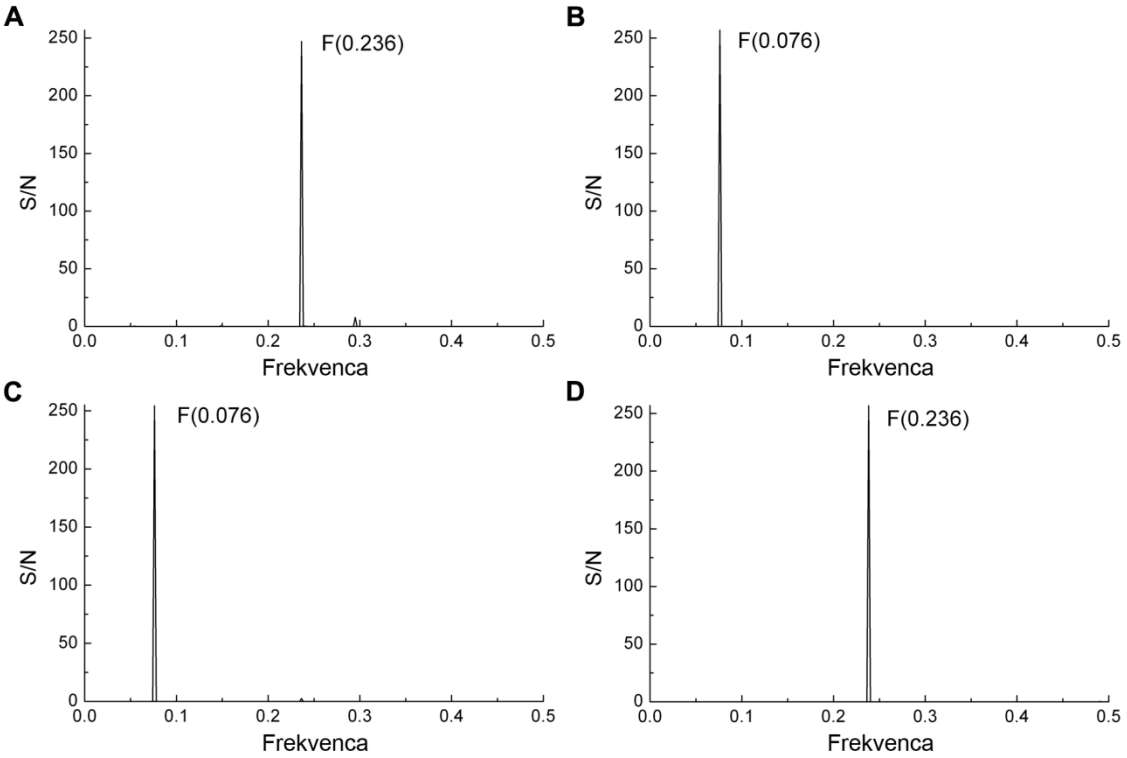
Slika 4.3.3.4. Upoređivanje osetljivosti standardnih filogenetskih metoda zasnovanih na MSA i ISM pristupima, na detekciju mutacija koje su bitne za humani tropizam egipatskog H5N1 virusa. (a),(c),(e) filogenetska stabla svih 309 nemutiranih G1 sekvenci; (b),(d),(f) stabla za sve G1 sekvence gde je svaka druga mutirana sa D43N, S120D, S129del, I151T. Stabla su kontruisana korišćenjem: (a),(b) standardne ML metode; (c),(d) standardne UPGMA metode; (e),(f) ISM algoritma. HA1 sekvence selektovane za mutacije su obojene tamno sivom bojom.



Slika 4.3.3.5. Graf po modelu elektro-opruga svih egipatskih H5N1 virusa izolovanih u Egiptu između 2006. i 2011. godine. Čvorovi grafa odnosno sekvence sa G1 specifičnim aminokiselinskim ostacima D43, S120, (S,L)129, I151 su obojene zelenom bojom, a sekvence sa G2 specifičnim ostacima N43, (D,N)120, 129del, T151 narandžastom bojom.

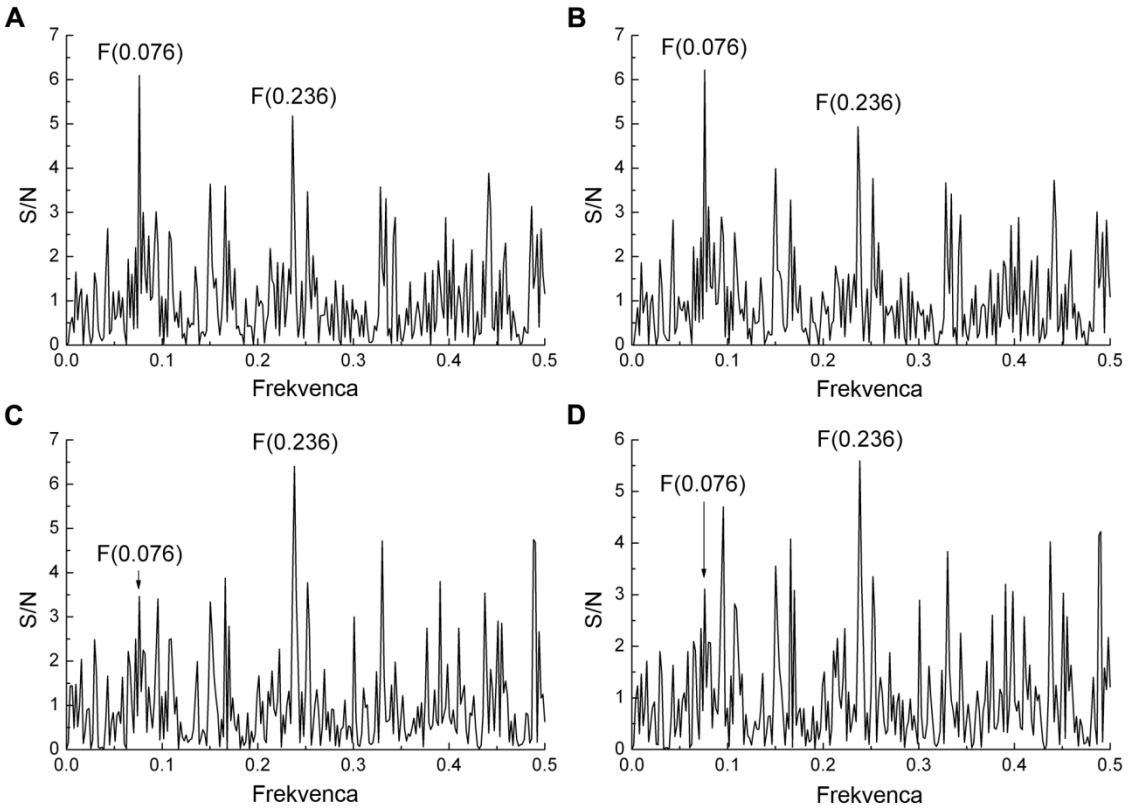
4.3.3.1.2. Analiza evolucije virusa H5N1 primenom ISM rastojanja odnosa amplituda A(0.236)/A(0.076)

Izračunat je informacioni spektar za 1576 objavljenih HA1 sekvenci sezonskog H1N1 virusa iz različitih godina i geografskih područja, dostavljenih u bazu između maja 2009 i novembra 2012. godine. Spektar reprezentativnog primera, kao i kros-spektar ove grupe virusa (slika 4.3.3.6a), pokazuje karakterističan pik na frekvenci F(0.236), što je u korelaciji sa rezultatima za manji skup sezonskog H1N1 virusa [224]. Kros-spektar svih 526 dostupnih H5N1-HPAIV HA1 sekvenci iz Egipta pokazuje karakterističan pik na frekvenci F(0.076) (slika 4.3.3.6b). Kros-spektar egipatskih sekvenci H5N1 HPAIV iz 2006-2008 poseduje pik samo na frekventnoj komponenti F(0.076) (slika 4.3.3.6c), dok kros-spektar sekvenci virusa iz 2009-2011 ima visoko dominantan pik na frekvenci F(0.236) (slika 4.3.3.6d). To pokazuje da egipatski H5N1-HPAIV ima istu karakterističnu frekvencu kao i sezonski H1N1 virusi.



Slika 4.3.3.6. Poređenje informacionih spektara sezonskog H1N1 i egipatskog H5N1 HPAIV. (a) Kros-spektar HA1 reprezentativnih sezonskih H1N1 virusa (n = 1576). (b) Kros-spektar svih trenutno dostupnih egipatskih H5N1 HPAIV (n = 526). (c) Kros-spektar egipatskih H5N1 HPAIV izolovanih između 2006-2008 i (d) 2009-2011.

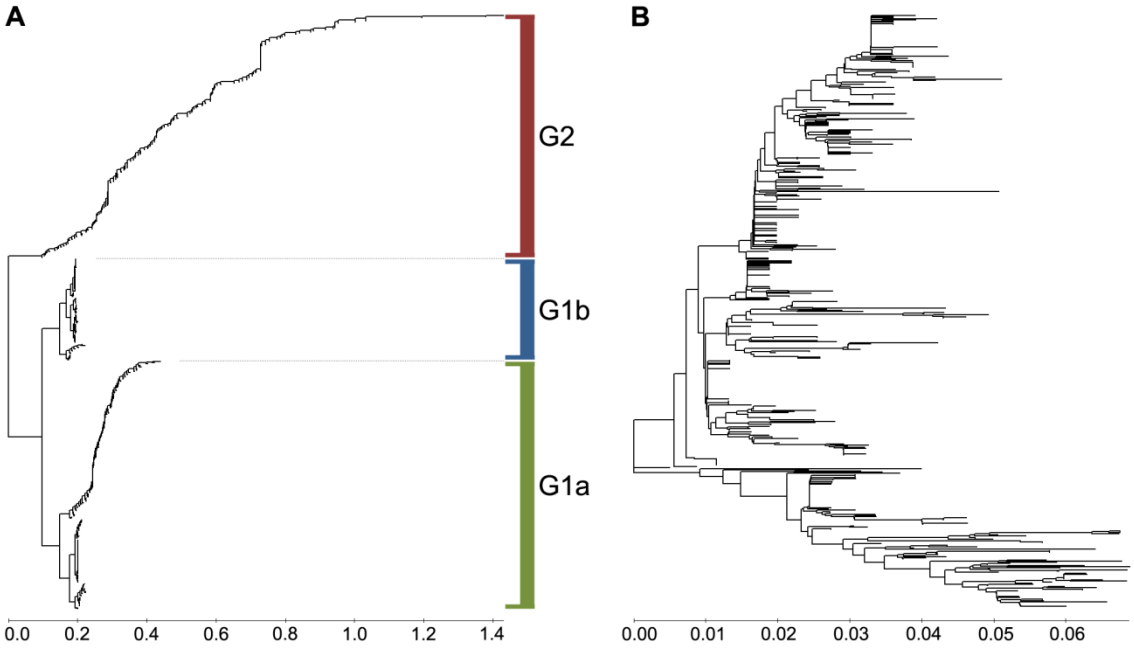
Slika 4.3.3.7 prikazuje pojedinačne informacione spektre egipatskog H5N1-HPAIV iz živine i ljudi u 2006 i 2010. godini. Informacioni spektri virusa iz 2006 (A/chicken/Egypt/R1/2006 (slika 4.3.3.7a) i A/Egypt/2763-NAMRU3/2006 (slika 4.3.3.7b)), imaju dominantni pik na frekvenci F(0.076) koji je tipičan za virus H5N1 ptičjeg receptora. H5N1 virusi izolovani u 2010 (A/chicken/Egypt/1029/2010 (slika 4.3.3.7c) i A/Egypt/N04434/2010 (slika 4.3.3.7d)) poseduju karakterističan pik na frekventnoj komponenti F(0.236) koja je tipična za sezonski H1N1 koji interreaguje samo sa humanim receptorom. Ovo sugeriše da su H5N1 virusi, koji kruže u Egiptu u 2010, evoluirali sa povećanjem afiniteta na humani receptor.



Slika 4.3.3.7. Poređenje informacionih spektara H5N1 HPAIV iz živine i ljudi izolovanih u periodu 2006-2010: (a) A/chicken/Egypt/R1/2006, (b) A/Egypt/2763-NAMRU3/2006, (c) A/chicken/Egypt/1029/2010, (d) A/Egypt/N04434/2010.

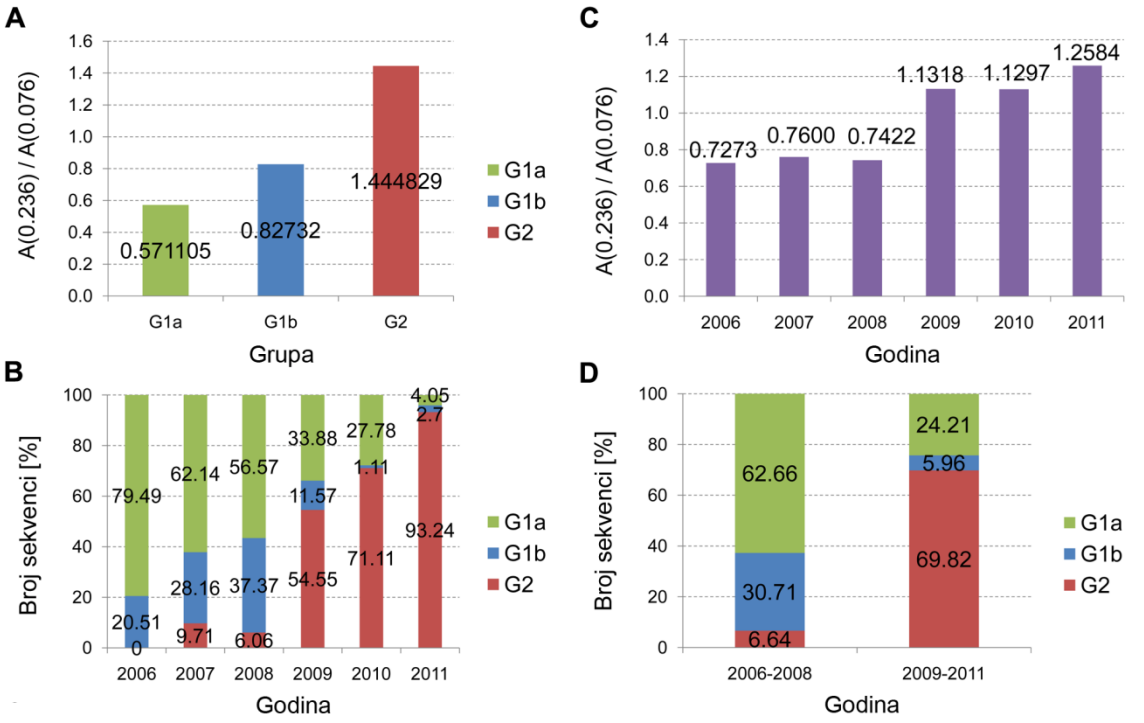
ISM filogenetsko stablo je konstruisano korišćenjem A(0.236)/A(0.076) odnosa amplituda kao rastojanja između H5N1-HPAIV HA1 sekvenci (slika 4.3.3.8a). Struktura stabla otkrila je pojavu tri odvojene grupe, G1A, G1B i G2. Sve H5N1

HPAIV sekvence u G1 imale su vrednost odnosa amplituda A(0.236)/A(0.076)<1 (prosečna vrednost 0.646±0.224, slika 4.3.3.9a), a svi virusi iz G2 imali su odnos A(0.236)/A(0.076)>1 (prosečna vrednost 1.445±0.311) (slika 4.3.3.9a).



Slika 4.3.3.8. MSA i ISM filogenetska analiza HPAIV H5N1 izolovanog u Egiptu. (a) Filogenetski stablo konstruisano koristeći filogenetsku metodu zasnovanu na ISM. (b) Filogenetski stablo konstruisano NJ metodom.

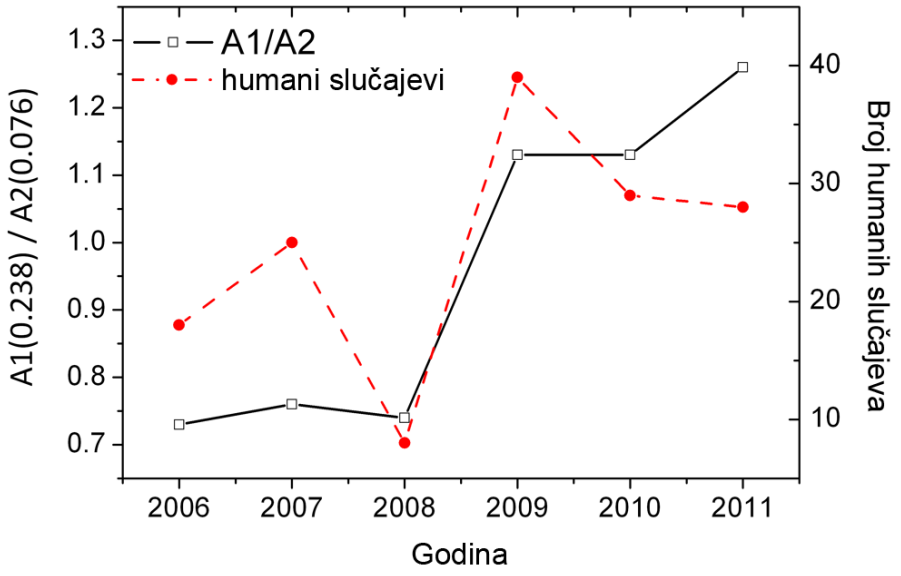
Struktura stabla takođe pokazuje dramatičan pomak od 100% G1 virusa u 2006, na 93,2% G2 virusa u 2011, uz značajno povećanje sa 6% na 54% u 2008/2009 (slika 4.3.3.9b). Nakon njihove najmasovnije pojave u 2008, G1b virusi praktično nestaju u 2009/2010 (slika 4.3.3.9b). Prosečne vrednosti A(0.236)/A(0.076) odnosa amplituda informacionih spektara svih egipatskih H5N1 HPAIV za svaku godinu, pokazali su značajan porast ovih vrednosti između 2006 i 2011 godine (slika 4.3.3.9c). Osim toga, 6,64% virusa sa A(0.236)/A(0.076)>1 su izolovani između 2006 i 2008, a 69,82% od tih virusa su izolovani u periodu 2009-2011. Dakle od 2006, sve više i više H5N1 HPAIV virusa u Egiptu je steklo karakterističnu IS funkciju (obrazac interakcija s receptorom) sezonskog H1N1.



Slika 4.3.3.9. Distribucija odnosa A(0.236)/A (0.076) po godinama za egipatske H5N1 HPAIV izolovane tokom 2006-2011. (a) Prosečne vrednosti A(0.236)/A(0.076) za G1 i G2 viruse. (b) Raspodela G1 i G2 virusa po godinama. (c) Prosečne vrednosti A(0.236)/A(0.076) po godinama. (d) Raspodela G1 i G2 virusa po trogodišnjim periodima.

Slika 4.3.3.10. pokazuje da je došlo do naglog povećanja broja humanih slučajeva nakon 2008, potvrđenih od strane Svetske zdravstvene organizacije (SZO) [266]. Zanimljivo je da je u periodu 2006-2008 bilo 14 puta više G1 virusa koji kruže (n = 225) od G2 virusa (n = 16), ali je kod ljudi pronađeno samo 4,1 puta više G1 virusa (n = 25) od G2 virusa (n = 6). U periodu 2009-2011, G2 virusa (n = 199) je bilo samo 2,3 puta više od virusa G1 (n = 86), ali su 9 puta češće zarazili ljude. Tako su tokom oba vremenska perioda G2 virusi inficirali ljude znatno češće nego što bi se očekivalo na osnovu njihove sveukupne prevalencije (kod ljudi i ptica).

Ovo sugeriše da virusi iz grupe G2 imaju povećani humani tropizam u poređenju sa virusima iz grupe G1. To dalje ukazuje da H5N1 HPAIV iz grupe G2 imaju veći afinitet za humani receptor i da su evoluirali u pravcu upotrebe humanog receptora, slično sezonskom H1N1 virusu.

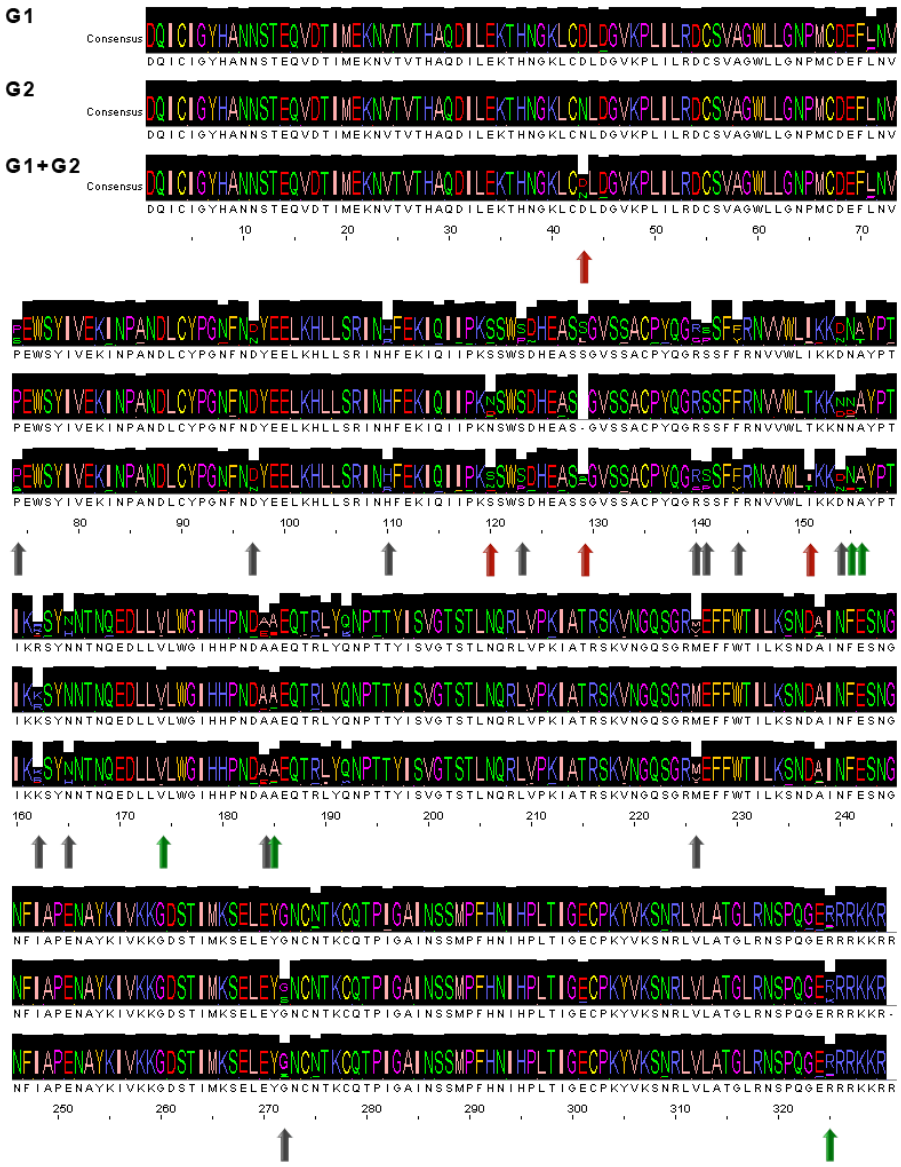


Slika 4.3.3.10. Korelacija između A(0.236)/A(0.076) vrednosti egipatskih H5N1 HPAIV i humanih slučajeva. Prikazane su prosečne vrednosti A(0.236)/A(0.076) po godinama i broj ljudskih slučajeva u Egiptu između 2006 i 2011. godine.

Analizom homologije 526 egipatskih H5N1 HPAIV HA1 sekvenci utvrđene su četiri grupe specifičnih aminokiselina (tabela 4.3.3.3, slika 4.3.3.11), koje se dosledno razlikuju između G1 i G2 virusa. Svi G1 virusi su imali aminokiseline: D43, S120, (S,L)129 i I151, dok je više od 94% od G2 virusa imalo: N43, (D, R) 120, 129del i T151 (aminokiselinska mesta su data po H5 numeraciji, tabela 4.3.3.4). Dok je u tabeli 4.3.3.4 navedeno još 18 pozicija u kojima se G1 i G2 virusi razlikuju, pozicije 43, 120, 129 i 151 su bile jedinstvene po njihovoj moći da naprave razliku između G2 i G1 sekvenci. Kako kod virusa živine, tako i kod humanih virusa, broj sekvenci sa G2-tipičnim aminokiselinama se stalno povećavao vremenom i posebno posle 2008 (tabela 4.3.3.5).

Tabela 4.3.3.3. Aminokiselinski ostaci koji su specifični za H5N1 HPAIV HA1 grupe G1 i G2 (slika 4.3.3.6a) i mutacije koje značajno povećavaju A(0.236)/A(0.076) odnos.

	Grupa G1	**Grupa G2**
Specifične aminokiseline	D43, S120, (S,L)129, I151	N43, (D,N)120, 129del, T151
Mutacije koje značajno povećavaju A(0.236)/A(0.076)	P74S, H110R, A127T, F143Y, K153D, S188K, S223(N,I), S234P, G272S, N275S	



Slika 4.3.3.11. Identifikacija nekonzerviranih pozicija u HA1 za G1 i G2 viruse: crvena/siva/zelena strelica predstavlja visok/srednji/mali procenat razlike.

Tabela 4.3.3.4. Raspodela specifičnih aminokiselina za grupe, po G1 i G2 virusima.

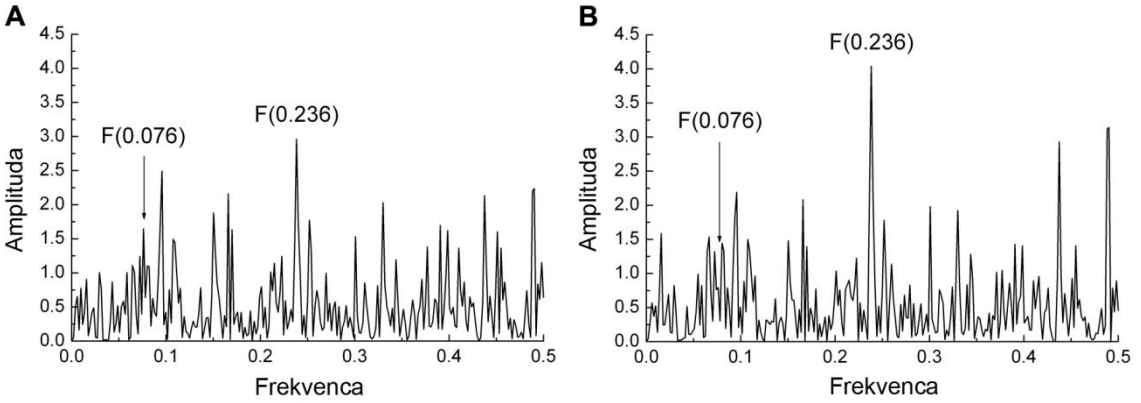
	Grupa G1	**Grupa G2**
Specifični aminokiselinski ostaci sa visokim procentom razlike > 90%	D43 (99%)	N43 (98%)
	S120 (94%)	(D,N)120 (34%,60%)
	(S,L)129 (71%,21%)	129del (99%)
	I151 (92%)	T151 (99%)
Specifični aminokiselinski ostaci sa srednjim procentom razlike 35% - 55%	S74 (40%)	P74 (100%)
	N97 (43%)	D97 (100%)
	R110 (39%)	H110 (100%)
	P123 (38%)	S123 (99%)
	G140 (35%)	R140 (98%)
	P141 (42%)	S141 (98%)
	Y144 (40%)	F144 (99%)
	H165 (39%)	N165 (100%)
	E184 (41%)	A184 (92%)
	(V,I)226 (39%,17%)	M226 (99%)
	G272 (99%)	S272 (40%)
Specifični aminokiselinski ostaci sa malim procentom razlike < 30%	D154 (65%)	N154 (61%)
	N155 (98%)	D155 (28%)
	T156 (25%)	A156 (98%)
	(I,E)162 (14%,10%)	(R,K)162 (43%,56%)
	V174 (97%)	I174 (5%)
	A185 (79%)	T185 (6%)
	R325 (81%)	K325 (26%)

Tabela 4.3.3.5. Broj aminokiselinskih supstitucija specifične za G2, koje su prikupljene pre i posle 2009. godine.

Mutacija	**Broj mutacija (2006-2008)**	**Broj mutacija (2009-2011)**
D43N	15 (7%)	200 (93%)
S120(N,D)	23 (12%)	196 (88%)
S129Δ	13 (6%)	191 (94%)
T151I	13 (6%)	207 (94%)

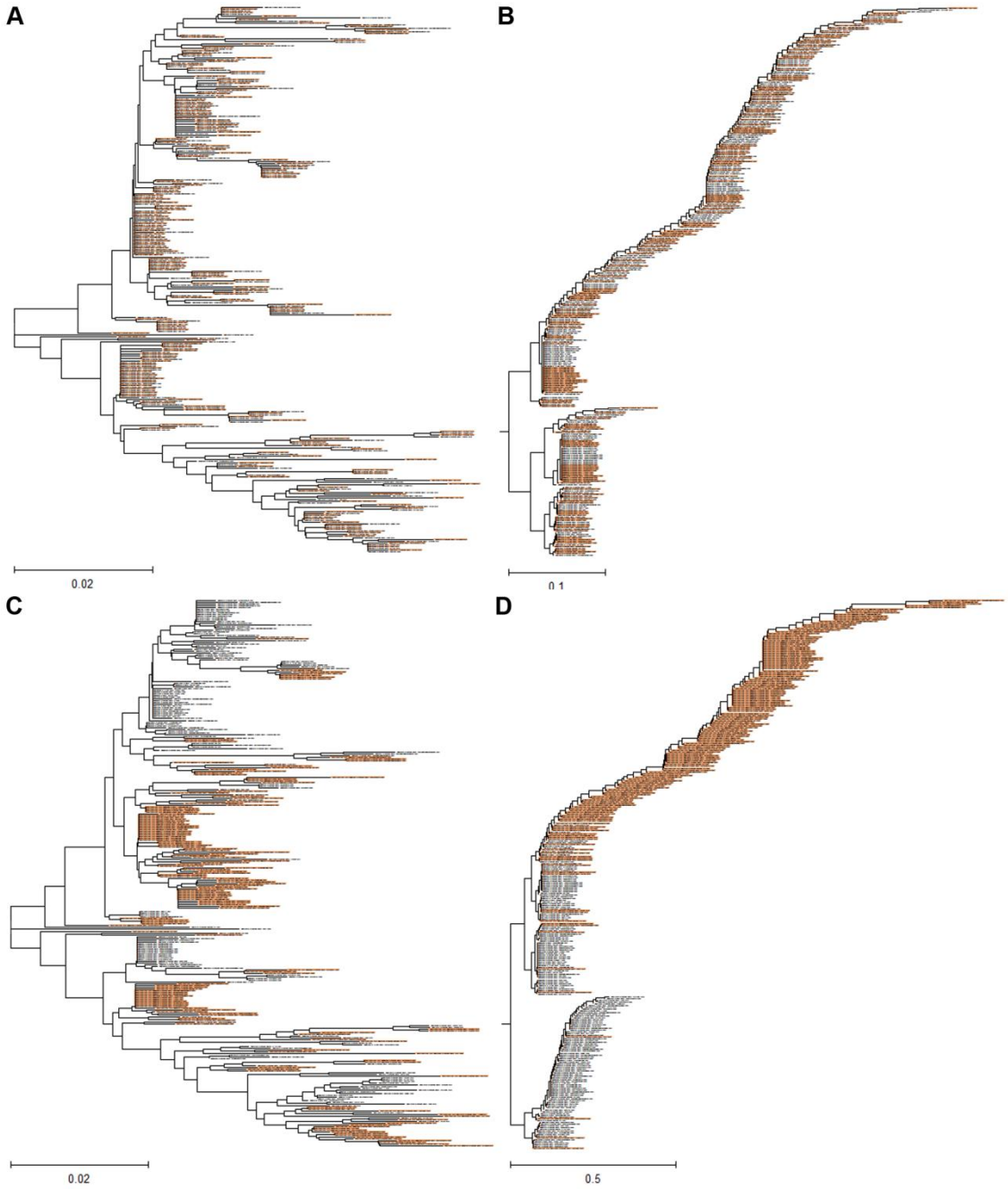
ISM filogenetsko stablo pokazuje da je u okviru grupe G2 i nakon sticanja promenjenog aminokiselinskog obrasca posle 2008, odnos A(0.236)/A(0.076) nastavio da se poveća sa 1,13 u 2009 do prosečne vrednosti 1,26 u 2011. Porast ovog odnosa je rezultat brojnih dodatnih mutacija koje su se nagomilavale vremenom. Tabela 4.3.3.3 pokazuje listu mutacija koje povećavaju vrednost odnosa A(0.236)/A(0.076), koje su pronađene u najmanje 2 soja analiziranih egipatskih virusa. Zanimljivo je, da je većina tih mutacija pronađena kod G1 virusa, ali G1 virusi su imali jednu, dve ili najviše tri mutacija iz tabele 4.3.3.3, što nije bilo dovoljno za povećanje odnosa A(0.236)/A(0.076) preko vrednosti 1.

U cilju procene kumulativnog efekta tih mutacija, upoređen je IS poslednje javno dostupne egipatske H5N1 HPAIV HA11 sekvence (A/Egypt/N04434/2010, prikupljena u martu 2010) (slika 4.3.3.12A), sa IS ovog istog HA1 sa svim mutacijama iz drugog reda u tabeli 4.3.3.3 (slika 4.3.3.12B). Kumulativni efekat ovih mutacija dramatično je povećao vrednost A(0.236)/A(0.076) od 1,79 kod prirodnog proteina (slika 4.3.3.12A) do 5,11 kod mutiranog proteina (slika 4.3.3.12B). Ovo sugeriše da je sticanje dodatnih mutacija (koje su već pronađene kod G1 virusa) kod G2 virusa prikupilo potencijal za dalje humano prilagođavanje, dok G1 virusi nisu imali taj potencijal. Dakle, aminokiselinske supstitucije date u tabeli 4.3.3.3 mogu predstavljati dragocene molekularne markera za dalje praćenje epidemioloških promena kod H5N1 HPAIV infekcija prema humanom prilagođavanju u Egiptu.



Slika 4.3.3.12. Efekat mutacija na informacione spektre egipatskog H5N1 HPAIV A/Egypt/N04434/2010. Efekat mutacija P74S, H110R, A127T, F143I, K153D, S188K, S223(N,I), S234P, G272S, N275S iz tabele 4.3.3.3, na IS HA1 H5N1 humanog virusa A/Egypt/N04434/2010. (a) IS nemutiranog HA1 proteina, (b) IS za HA1 sa mutacijama.

Slika 4.3.3.8b pokazuje filogenetsko stablo za egipatske HPAIV H5N1, na osnovu NJ i ML metoda. Oba stabla imaju veoma sličnu strukturu potvrđenu visokim *bootstrap* vrednostima [267]. Slično kao kod ISM stabla (slika 4.3.3.8A), većina G2 virusa pojavljuju se u zajedničkoj podgrani konvencionalnih stabala. Upoređena je osetljivost filogenetskog pristupa zasnovanog na MSA, sa ISM filogenetskim pristupom, na identifikaciju funkcionalno značajnih mutacija. Slike 4.3.3.13a i b predstavljaju MSA i ISM stabla HA1 sekvenci svih 311 G1 virusa. Uvedene su četiri aminokiseline (43N, 120D, 129del, 151T) karakteristične za G2 virusa u svaku drugu sekvencu izabranog G1 skupa. Slike 4.3.3.13c i d pokazuju da je za razliku od MSA stabla, koje nije osetljivo na mutacije, ISM stablo jasno odvaja mutirane sekvence u G2 klaster.



Slika 4.3.3.13. Poređenje osetljivosti MSA i ISM filogenetskog stabla na mutacije. Mutacije koje su potencijalno važne za humani tropizam egipatskog H5N1 HPAIV su uključeni u test osetljivosti MSA i ISM filogenetskih pristupa na mutacije. (a) MSA stablo (b) ISM stablo za svih 311 nemutiranih H5N1 HPAIV HA1 sekvenci G1 virusa, (c) MSA stablo panela (a) u kojoj je svaka druga sekvenca mutirana i (d) ISM stablo koje odgovara mutiranim i nemutiranim sekvencama prikazanih na panelu (c). H5N1 HPAIV HA1 sekvence odabrane za mutacije D43N, S120D, S129del i I151T su označene narandžastom bojom.

Rezultati ISM filogenetske analize egipatskih H5N1 virusa pokazuju: (i) grupisanje H5N1 na dve odvojene grupe G1 i G2, (ii) konstantan rast od 2006 do 2011 u broju G2 virusa, što je u korelaciji sa povećanjem broja humanih slučajeva influence, (iii) identifikaciju karakterističnih aminokiselinskih pozicija D43, S120, (S,L)129, I151 u G1, i N43, (D,N)120, 129del, T151 u G2 i (iv) karakterisanje četiri G2 mutacije kao bitnih za povećanje humanog tropizma H5N1.

Pored toga, ISM filogenetska analiza je pokazala prednost u odnosu na standardne metode u smislu: (i) osetljivosti na mutacije i delecije koje su bitne za biološku funkciju proteina i (ii) osetljivosti na poziciju i vrstu pojedinačne aminokiselinske supstitucije.

Upoređivanje rezultata ISM filogenetske analize H5N1 virusa primenom rastojanja na celom spektru i koristeći ISM rastojanje odnosa amplituda A(0.236)/A(0.076) pokazuje: (i) generisanje sličnih filogenetskih stabala, sa istim klasterisanjem na dve grupe G1 i G2, (ii) identifikaciju istih karakterističnih aminokiselinskih supstitucija u G1 i G2, u oba slučaja (tabele 4.3.3.2 i 4.3.3.3).

4.3.4. Selekcija terapeutskih malih molekula

Razvoj leka je kompleksan, skup i dugotrajan proces. Virtuelni skrining (VS), *in silico* pristup, kojim se izdvajaju kandidati za lekove od velikog broja jedinjenja iz molekulskih biblioteka, donosi uštedu od oko 130 miliona dolara i 0,8 godina po jednom leku [268].

Za virtuelni skrining, baziran na molekulskim deskriptorima AQVN i EIIP, koji je primenjen je na selekciju anti-HIV jedinjenja i antibiotika, korišćeni su programi EIIP/ISM platforme: Chemdb2Alati, FormulaKalkulator, PubchemParser i Raspodela2D (Modul za pretraživanje molekulskih biblioteka).

4.3.4.1. Selekcija terapeutskih malih molekula u terapiji HIV-a

Procenjuje se da je oko 33 miliona ljudi inficirano HIVom, koji predstavlja jedan od deset glavnih uzroka mortaliteta u svetu [269]. Uvođenje efikasne antriretroviralne terapije (eng. *highly active antiretroviral therapy, HAART*), za koju su biološki targeti reverzna transkriptaza i proteaza (proteini virusa HIV-a), je značajno uticalo na produženje i kvalitet života HIV bolesnika [270]. Iako je HAART terapija vrlo efikasna, pojava rezistencije na lek, latentni virusni rezervoari, i toksičnost terapije ukazuju na potrebu razvoja novih lekova u terapiji HIV-a [271].

Rezultati su objavljeni u sledećim radovima [272, 273, 136]:
*Maga G, Veljkovic N, Crespan E, Spadari S, Prljic J, **Perovic V**, Glisic S, Veljkovic V. New in silico and conventional in vitro approaches to advance HIV drug discovery and design. Expert Opin Drug Discov. 2013 Jan;8(1):83-92. doi: 10.1517/17460441.2013.741118. Epub 2012 Nov 20. Review. PubMed PMID: 23167743.*
*Veljkovic N, Glisic S, Prljic J, **Perovic V**, Veljkovic V. Simple and General Criterion for "In Silico" Screening of Candidate HIV Drugs. Curr Pharm Biotechnol. 2012 Mar 20. [Epub ahead of print] PubMed PMID: 22429138.*
*Tintori C, Manetti F, Veljkovic N, **Perovic V**, Vercammen J, Hayes S, Massa S, Witvrouw M, Debyser Z, Veljkovic V, Botta M. Novel virtual screening protocol based on the combined use of molecular modeling and electron-ion interaction potential techniques to design HIV-1 integrase inhibitors. J Chem Inf Model. 2007 Jul-Aug;47(4):1536-44.*

U selekciji terapeutskih malih molekula u terapiji HIV-a, primenjen je EIIP/AQVN kriterijum za brz i efikasan *in silico* skrining molekulskih biblioteka u cilju pronalaženja jedinjenja koja poseduju anti-HIV1 aktivnost. Selekcija malih molekula u odnosu na njihove EIIP/AQVN karakteristike omogućava smanjenje broja kandidata za lekove u svakoj od klasa anti HIV lekova - CxCR4 i CCR5 inhibitora, integraznih inhibitora (INI), proteaznih inhibitora (PI), nukleotidnih (NtRTI) i nukleozidnih (NRTI) inhibitora reverzne transkriptaze (RT) i anti-HIV flavonoida.

VS AQVN/EIIP kriterijum će prvo biti pokazan na primeru PI. Tabela 4.3.4.1 sadrži trening set sa 23 PI. 89% od odobrenih PI ima AQVN u intervalu 2.61–2.78 i EIIP vrednosti u intervalu 0.04–0.08 Ry. Ovaj EIIP/AQVN opseg je odabran za selekciju kandidata PI. Zatim je kriterijum za selekciju validiran na jedinjenjima sa PI aktivnošću iz HIV/OI terapeutske baze [156] i ustanovljeno je da 89.4% analiziranih jedinjenja zadovoljava ovaj kriterijum. Kao negativna kontrola korišćeni su CCR5 inhibitori iz iste anlizirane antiHIV molekulske biblioteke. Samo 27,9% (247 od 885) CCR5 inhibitora je bilo u opsegu EIIP/AQVN koja karakteriše jedinjenja sa PI aktivnošću. Ovaj rezultat potvrđuje veliku selektivnost EIIP/AQVN kriterijuma u selekciji kandidata za PI.

Tabela 4.3.4.1. Trening set za HIV1 PI

Proteazni inhibitor	**Formula**	**AQVN**	**EIIP [Ry]**
Amprenavir	C25H35N3O6S	2.743	0.0488
Atazanavir	C38H52N6O7	2.680	0.0662
Indinavir	C36H47N5O4	2.610	0.0814
Lopinavir	C37H48N4O5	2.617	0.0799
Nelfinavir	C32H45N3O4S	2.567	0.0985
Ritonavir	C37H48N6O5S2	2.735	0.0512
Saquinavir	C38H50N6O5	2.646	0.0739
Tipranavir	C31H33F3N2O5S	2.747	0.0476
Darunavir	C27H37N3O7S	2.773	0.0392
PL-100	C33H44N4O6S	2.704	0.0597
DMP323	C35H38N2O5	2.725	0.0540
TMC 126	C28H38N2O8S	2.779	0.0373
Doxovir	C20H30N6O2	2.621	0.0792
DPC-681	C35H48FN5O5S	2.632	0.0770
DPC-684	C35H48FN5O5S	2.632	0.0770
L-756, 423	C39H48N4O5	2.646	0.0740
Brecanavir	C33H41N3O10S2	2.921	0.0139
Mozenavir	C33H36N4O3	2.710	0.0581

RO033-4649	C37H55N5O7S	2.629	0.0776
PD-178390	C28H37NO4S	2.592	0.0843
LB-71350	C33H45N3O8S	2.733	0.0516
AG-1776	C32H37N3O5S	2.769	0.0406
UIC02031	C30H39N3O8S	2.815	0.0253

Zatim je vršena selekcija inhibitora HIV1 integraze (INI) koja je počela analizom 1956 jedinjenja iz HIV/OI terapeutske baze [156]. Ovom analizom ustanovljeno je da inhibitore HIV1 integraze karakteriše AQVN interval 3.00-3.20 i EIIP opseg 0.044-0.116 [136]. Navedeni AQVN/EIIP domeni su potom korišćeni prilikom virtualnog skrininga (VS) 200000 jedinjenja iz *Asinex Gold Colection* baze kandidata HIV integraznih inhibitora [159]. Ovom analizom je izdvojeno 9600 jedinjenja (4,8% od analiziranih jedinjenja), koja su zatim bila analizirana primenom VS koji se bazira na strukturi (rotirajuće veze, modeli farmakofora, doking). Posle ove analize 12 jedinjenja je izdvojeno za testiranje biološkim esejima. Finalno je jedan kandidat za HIV1 integrazni inhibitor selektovan za pretkliničku studiju [274].

Ranijim analizama su određeni AQVN/EIIP intervali koji karakterišu HIV1 entri inhibitore CCR5 i flavonoide sa anti HIV aktivnošću: AQVN interval 2.42-2.63 i EIIP opseg 0.079-0.099Ry za CCR5, i AQVN domen 3.34-3.59 i EIIP interval 0.110-0.135Ry za anti-HIV flavonoide [275, 135].

Hemokinski receptor 4 (CXCR4, takođe poznat kao fuzin) je protein koji se nalazi na površini nekih imunskih ćelija. To je jedan od dva koreceptora za HIV pored CD4 receptora za koji se HIV vezuje i ulazi u ćeliju domaćina. Subtipovi HIV, za koje je CXCR4 koreceptor, su patogeniji i pojavljuju se kasnije u HIV infekciji i korelišu sa smanjenjem broja CD4 ćelija i brzom progresijom bolesti. Pošto je trenutno mali broj anti-CXCR4 lekova u pretkliničkoj i kliničkoj fazi razvoja leka, nije bilo moguće formirati pogodan set za treniranje. Stoga su preuzeta jedinjenja iz HIV/OI terapeutske baze [156] klasirana kao CXCR4 inhibitori, i određeni su domeni AQVN i EIIP koji obuhvataju 80% ovih molekula. Od 105 jedinjenja iz HIV/OI terapeutske baze 85 (81%) ima AQVN u intervalu 2.16–2.53 i EIIP vrednosti u intervalu 0.062–0.096 Ry. Ovaj opseg AQVN i EIIP je uzet kao kriterijum za selekciju kandidata za CXCR4 inhibitore. Analiza CCR5 inhibitora iz anti-HIV baze [156] kao negativna kontrola je pokazala da

338 (38.2%) ovih jedinjenja ima AQVN i EIIP vrednosti u opsegu koji karakteriše jedinjenja sa anti-CXCR4 aktivnošću. Istovremeno, nijedan od molekula iz PI i NRTI/NtRTI trening seta ne ispunjava kriterijum za CXCR4 inhibitore. Ovaj rezultat ukazuje da AQVN opseg 2.16–2.53 i EIIP interval 0.062–0.096 Ry predstavlja dobar kriterijum za selekciju kandidata za CXCR4 inhibitore.

Nukleotidni (NtRTI) i nukleozidni (NRTI) inhibitori reverzne transkriptaze, NtRTIs zajedno sa ne-nukleozidnim inhibitorima RT predstavljaju glavnu komponentu HAART terapije. Definisan je EIIP/AQVN kriterijum za VS molekularnih biblioteka na sledeći način. U tabeli 4.3.4.2. je prikazan trening set kojeg čine 32 jedinjenja sa anti RT aktivnošću [268]. Od toga 28 jedinjenja (87.5%) imaju AQVN vrednosti 2.92– 3.20 i EIIP u intervalu 0.04–0.10 Ry. Ovaj interval je izabran kao kriterijum za selekciju kandidata NtRTI i i NRTI VS. Kao negativna kontrola korišćeni su CCR5 inhibitori iz iste antiHIV molekulske biblioteke. Samo 2.9% (26 od 885) CCR5 inhibitora, zajedno sa jednim od 23 PI iz tabele 4.3.4.1, su u opsegu EIIP/AQVN koja karakteriše jedinjenja sa NtRTI i i NRTI aktivnošću. Ovaj rezultat potvrđuje veliku selektivnost EIIP/AQVN kriterijuma u selekciji kandidata za NtRTI i NRTI inhibitore.

Tabela 4.3.4.2. Trening set za inhibitore HIV-1 RT.

NRTI/NtRTI	**Formula**	**AQVN**	**EIIP [Ry]**
Abacavir	C214H18N6O	2.820	0.0233
Didanosine	C10H12N4O3	3.103	0.0810
Emtricitabine	C8H10FN3O3S	3.154	0.0967
Lamivudine	C8H11N3O3S	3.154	0.0967
Stavudine	C10H12N2O4	3.071	0.0700
Tenofovir	C9H14N5O4P	3.152	0.0960
Zalcitabine	C9H13N3O3	2.930	0.0167
Zidovudine	C10H13N5O4	3.188	0.1060
Apricitabine	C8H11N3O3S	3.154	0.0967
Racivir	C8H10FN3O3S	3.154	0.0967
MIV-210	C10H12FN5O3	3.097	0.0787
Elvucitabine	C9H10FN3O3	3.077	0.0720

Dioxolane Thymidine	C9H12N2O5	3.143	0.0934
AVX754	C8H11N3O3S	3.154	0.0967
KP-1461	C8H14N4O4	3.000	0.0439
Beta-fluoro-DDA	C10H12FN5O2	3.000	0.0439
Emivirine	C17H22N2O3	2.682	0.0656
Lobucavir	C11H15N5O3	3.000	0.0439
SPD756	C12H16N6O3	3.027	0.0540
dOTC	C8H11N3O3S	3.154	0.0967
FddA	C10H12FN5O2	3.000	0.0439
Alovudine	C10H13FN2O4	2.933	0.0185
Dexelvucitabine	C9H10FN3O3	3.077	0.0720
Amdoxovir	C9H12N6O3	3.200	0.1092
Adefovir Dipivoxil	C20H32N5O8P	2.879	0.0022
BEA-005	C10H15N3O4	2.939	0.0201
BETA-D-FD4C	C9H10FN3O3	3.077	0.0720
Phosphasid	C10H14N5O6P	3.333	0.1319
(+)-.BETA.-D-FDOC	C8H10FN3O4	3.154	0.0967
DTTP	C10H17N2O14P3	3.608	0.1006
Stampidine	C20H23BrN3O8P	3.071	0.0700
4'-Ethynyl D4T	C12H12N2O4	3.133	0.0905

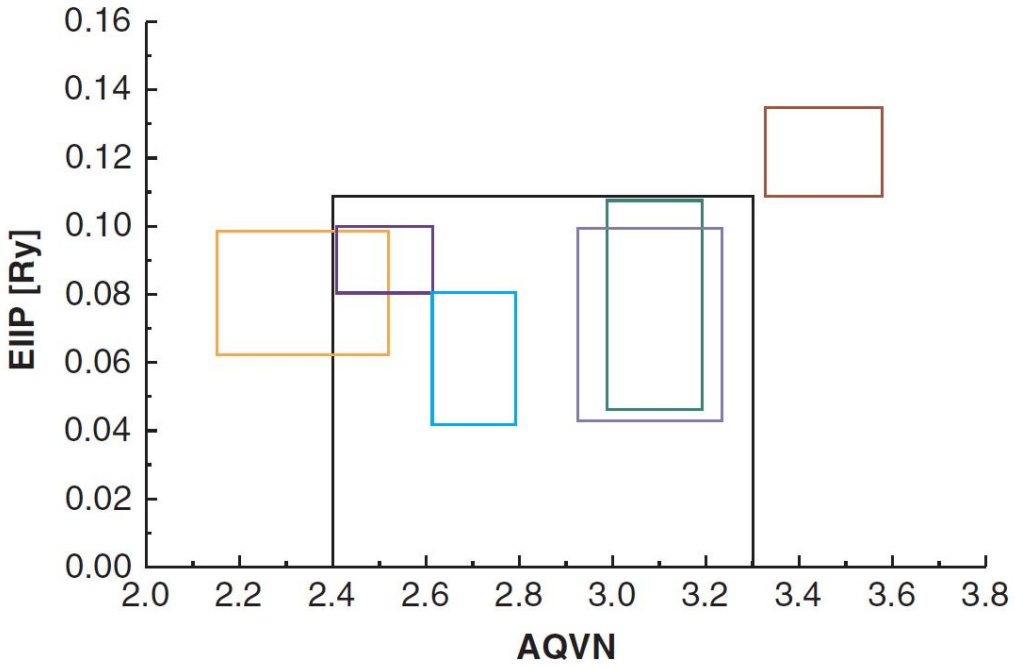
Prikazanom analizom anti HIV lekova ustanovljen je opšti EIIP/AQVN kriterijum za VS, i to za svaku od navedenih klasa anti HIV lekova: CxCR4 i CCR5 inhibitora, integraznih inhibitora (INI), proteaznih inhibitora (PI), nukleotidnih (NtRTI) i nukleozidnih (NRTI) inhibitora reverzne transkriptaze i anti-HIV flavonoida, predstavljen u tabeli 4.3.4.3. i na slici 4.3.4.1.

Treba napomenuti da se AQVN/EIIP domeni za CXCR4 i CCR5 entri inhibitore preklapaju, kao i za NRTI/NtRTI i INI, ukazujući da neki od terapeutskih molekula može pokazati dvostruku anti-HIV aktivnost. To je u saglasnosti sa podacima Wanga i saradnika, koji su otkrili da neka jedinjenja istovremeno inhibiraju RT u malim nanomolarnimi i IN u malim mikromolarnim koncentracijama [276]. Ovo istraživanje,

kao i rezultati ovog rada sugerišu da je potrebno usmeriti napor u cilju pronalaženja jednog leka koji bi inhibirao oba navedena enzima, pošto bi terapija koja bi delovala na više targeta istovremeno, mogla imati manju toksičnost, jednostavnije doziranje i lakšu primenu kod pacijenata.

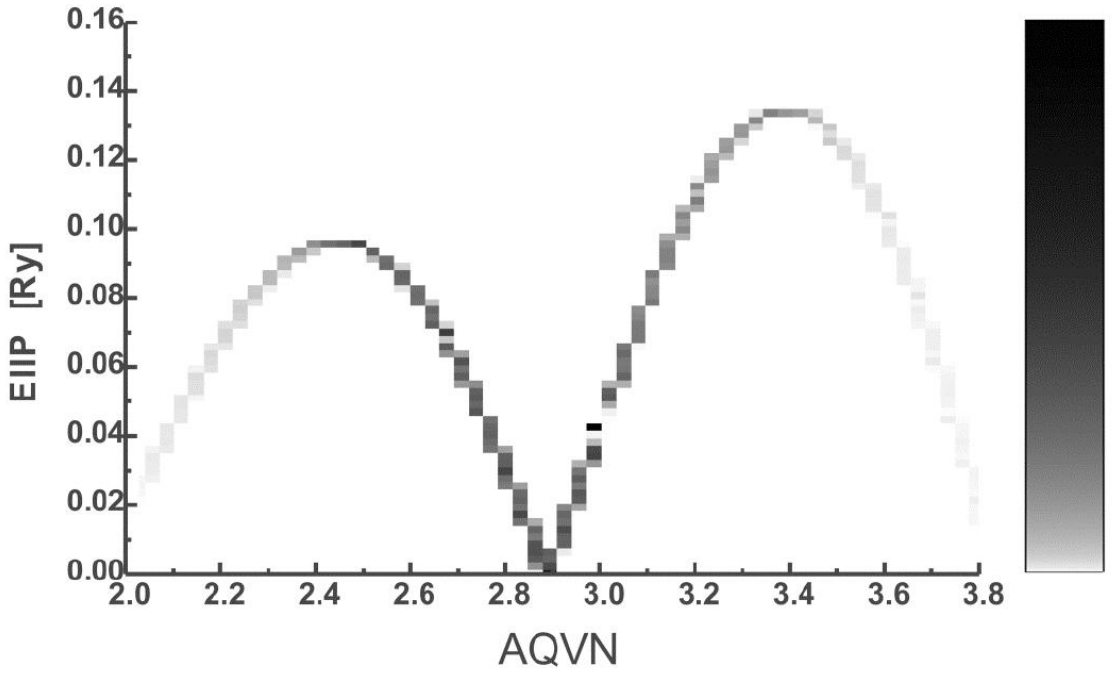
Tabela 4.3.4.3. EIIP/AQVN kriterijum za VS molekulskih biblioteka u selekciji malih anti-HIV molekula.

Target	**AQVN**	**EIIP [Ry]**
CXCR4	2.16 – 2.53	0.062 – 0.096
CCR5	2.42 – 2.63	0.079 – 0.099
PI	2.61 – 2.78	0.040 – 0.080
NRTI / NtRTI	2.92 – 3.20	0.040 – 0.100
INI	3.00 – 3.20	0.044 – 0.116
Anti-HIV flavonoidi	3.34 – 3.59	0.110 – 0.135

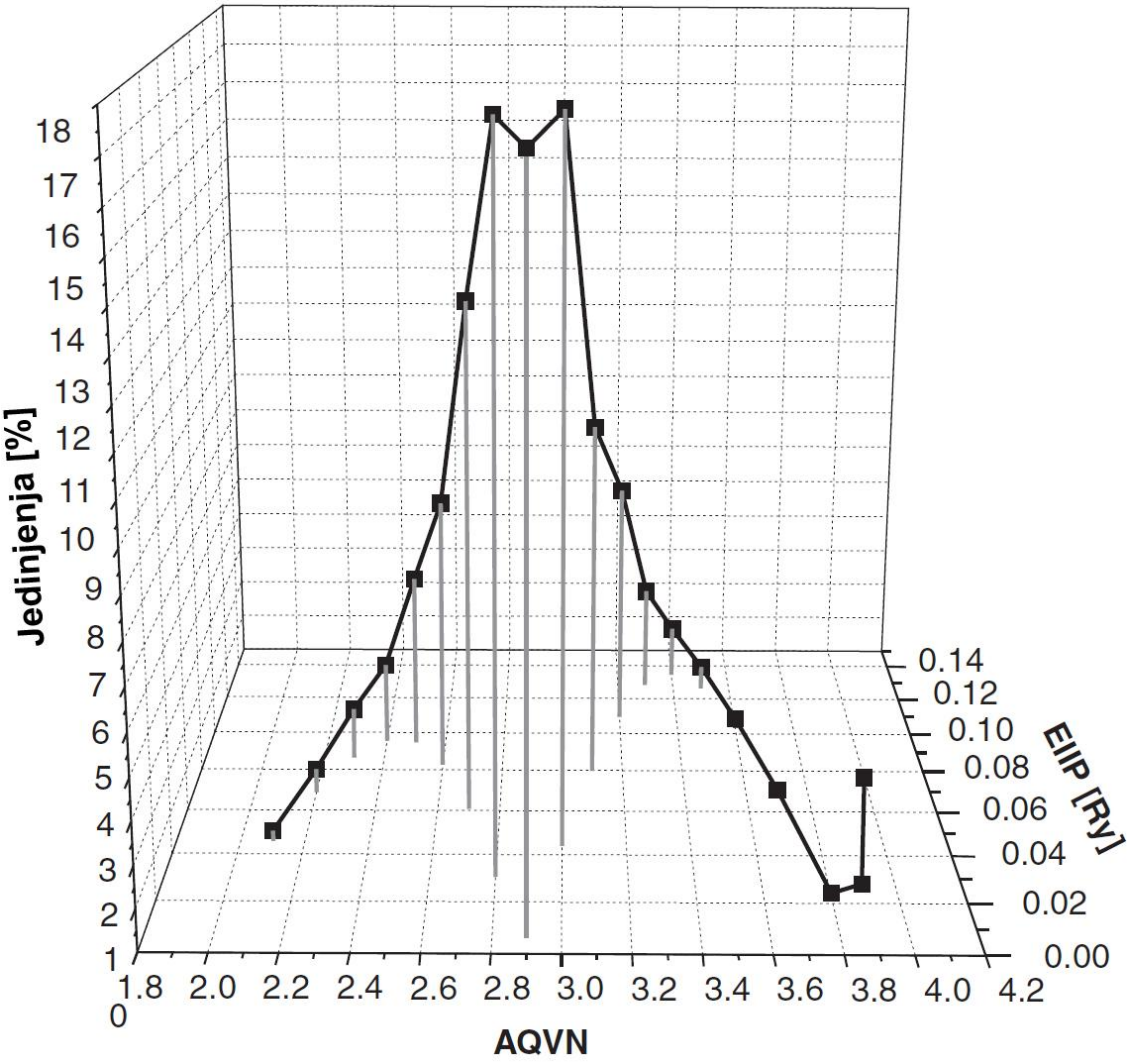


Slika 4.3.4.1. Šematska prezentacija EIIP/AQVN kriterijuma za selekciju kandidata anti-HIV jedinjenja VS molekulskih biblioteka: žuto - CXCR4 inhibitori, ljubičasto - CCR5 inhibitori, tirkiz - PI, plavo - NRTI/NtRTI, zeleno - INI, bordo - anti-HIV flavonoidi; crno - EIIP/AQVN domen homologe distribucije za više od 90% jedinjenja iz PubChem baze.

U cilju ispitivanja specifičnosti predloženog EIIP/AQVN kriterijuma za selekciju anti-HIV jedinjenja, poređene su AQVN i EIIP vrednosti anti-HIV jedinjenja sa vrednostima ovih molekulskih deskriptora za 45010644 jedinjenja iz PubChem baze [157]. Utvrđeno je da većina jedinjenja iz PubChem baze (92.5%) ima homogenu distribuciju na EIIP intervalu 0.00-0.11 Ry i AQVN intervalu 2.4-3.3 (slika 4.3.4.1). Ovaj domen EIIP/AQVN prostora, koji obuhvata najveći deo poznatih hemijskih jedinjenja, je označen kao „osnovni EIIP/AQVN hemijski prostor“ (eng. *basic chemical space, BCS*). Inhibitori CCR5 obuhvataju 7.2%, PI 6.9%, NtRTI/NRTI 14.5% i INI 17.0% BCS prostora. Svi anti-HIV flavonoidi i većina CXCR4 inhibitora se nalazi izvan BCS prostora.



Slika 4.3.4.2. Distribucija jedinjenja iz PubChem baze prema njihovim EIIP I AQVN deskriptorima. Podaci o broju jedinjenja sa određenim AQVN i EIIP vrednostima su označeni prema sivoj skali u desnoj koloni: crno označava 10000 jedinjenja, belo 0, sivo srednji broj jedinjenja.



Slika 4.3.4.3. Distribucija 45010644 jedinjenja iz PubChem baze u odnosu na AQVN i EIIP vrednosti.

4.3.4.2. Selekcija antibiotika

Mortalitet i morbiditet uzrokovan patogenim bakterijama rezistentnim na više klasa antibiotika (eng. *multi drug resistant, MDR*) je u stalnom porastu. To ukazuje na potrebu za otkrićem novih antibiotika za tretman MDR.

Rezultati su objavljeni u sledećem radu [277]:
*Veljkovic N, Glisic S, **Perovic V**, Veljkovic V. The role of long-range intermolecular interactions in discovery of new drugs. Expert Opin Drug Discov. 2011 Dec;6(12):1263-70. doi: 10.1517/17460441.2012.638280. Epub 2011 Nov 15. PubMed PMID: 22647065.*

Analizom 230 antibiotika, koji spadaju u različite antibiotske klase, utvrđen je AQVN/EIIP opseg za sledeće antibiotske klase: peniciline, cefalosporine, karbapeneme i peneme, monobaktame, hinoline, aminoglikozide, tetracikline, makrolide, pleuromutiline, nitrofurane (tabela 4.3.4.4).

Tabela 4.3.4.4. Domeni AQVN i EIIP koji karakterišu određene antibiotske klase.

Klasa antibiotika	**AQVN**	**EIIP [Ry]**
Penicilini	2.975 - 3.180	0.035 - 0.124
Cefalosporini	3.071 - 3.473	0.070 - 0.130
Karbapenemi i Penemi	2.973 - 3.059	0.022 - 0.066
Monobaktami	3.166 - 3.581	0.100 - 0.134
Hinolini	2.760 - 3.060	0.003 - 0.065
Aminoglikozidi	2.552 - 2.820	0.024 - 0.084
Tetraciklini	2.933 - 3.111	0.018 - 0.084
Makrolidi	2.467 - 2.630	0.077 - 0.096
Pleuromutilini	2.395 - 2.473	0.095 - 0.096
Nitrofurani	3.652 - 3.826	0.010 - 0.086

MDR Enterobacteriaceae su bakterije koje sadrže enzim metalo beta laktamazu (NDM1) koji čini ove bakterije rezistentnim na sve beta laktame uključujući i karbapeneme koji se smatraju zadnjom linijom odbrane od MDR bakterija. Analizom antibiotika na koje je MDR Enterobacteriaceae rezistentna [278], utvrđen je opseg rezistentnosti kojeg karakterišu intervali AQVN 2.55-3.42 i EIIP 0.00-0.13 Ry. Na osnovu podataka iz tabele 4.3.4.4, može se zaključiti da ovi antibiotici pokrivaju AQVN/EIIP prostor skoro svih klasa antibiotika sa izuzetkom pleuromutilina i

nitrofurana. Ovaj rezultat ukazuje da ove dve klase antibiotika treba testirati protiv NDM 1 pozitivnih bakterija.

Prikazani rezultati ukazuju da primena bioinformatičkih tehnika koje su bazirane na konceptu dalekodosežnih interakcija, uporedo sa ostalim računarskim tehnikama, može da unapredi i ubrza otkriće novih lekova.

5. Zaključak

Značaj istraživanja i naučni doprinos

U skladu sa ciljevima rada, razvijena je multifunkcionalna platforma za analizu bioloških molekula i njihove međusobne interakcije. Primena ove platforme je pokazala da dalekodosežne međumolekulske interakcije, determinisane potencijalom elektron-jon interakcije (EIIP) i srednjim valentnim brojem (AQVN), igraju značajnu ulogu u biološkim sistemima i da predstavljaju osnov za razvoj bioinformatičkih i hemoinformatičkih pristupa koji mogu imati široku primenu u molekularnoj biologiji i biomedicini. Osnovni doprinosi ovog rada mogu se sumirati u sledećem:

Razvijena je originalna EIIP/ISM platforma koja uključuje sledeće osnovne module:

(a) Modul za određivanje spektralnih karakteristika koje reprezentuju informaciju kodiranu u primarnim strukturama proteina, koja determiniše njihovu biološku funkciju.

(b) Modul za identifikaciju proteina čije primarne strukture kodiraju komplementarnu informaciju koja omogućava njihovu međusobnu interakciju.

(c) Modul za određivanje domena primarne strukture proteina koji daju dominantan doprinos informaciji odgovornoj za dalekodosežnu protein–protein interakciju.

(d) Modul za analizu uticaja mutacija na karakteristike informacionog spektra koji omogućava procenu njihovog delovanja na biološku funkciju proteina.

(e) Modul za podešavanje informacionih spektara promenama u primarnoj strukturi (mutacije, delecije, insercije) koji omogućava modulaciju biološke funkcije proteina i "de novo" dizajn peptida željene biološke aktivnosti.

(f) Modul za filogenetsku analizu proteina koji omogućava procenu doprinosa pojedinačnih mutacija funkcionalnoj evoluciji proteina.

(g) Modul za predviđanje biološke funkcije malih molekula koji omogućava pretraživanje molekulskih biblioteka u cilju selekcije jedinjenja koja predstavljaju kandidate za nove lekove.

(h) Pomoćni moduli koji omogućavaju manipulisanje molekulskim bazama podataka koje su neophodne za rad osnovnih modula.

Primena EIIP/ISM platforme u rešavanju različitih problema kao što su: određivanje terapeutskih targeta za razvoj lekova i vakcina za grip, identifikacija mutacija na LPL proteinu koje predstavljaju faktor rizika za kardiovaskularne bolesti, analiza funkcionalne evolucije virusa gripa i procena njihovog pandemijskog potencijala, selekcija malih molekula koji predstavljaju kandidate za nove lekove za SIDU i nove antibiotike, pokazala je široku praktičnu primenjivost ove platforme.

Na osnovu svega izloženog, može se zaključiti da EIIP/ISM platforma predstavlja nov i, zbog svoje biofizičke osnove, jedinstven bioinformatički i hemoinformatički softverski paket za rešavanje širokog spektra problema u molekularnoj biologiji i biomedicini.

Literatura

[1] GenBank Release Notes. URL http://www.ncbi.nlm.nih.gov/genbank/statistics.

[2] UniProtKB/Swiss-Prot Release 2013 statistics. URL http://web.expasy.org/docs/relnotes/relstat.html.

[3] PubChem Help. URL http://pubchem.ncbi.nlm.nih.gov/help.html.

[4] Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. Methods of information in medicine, 40(4), 346-358.

[5] Bioinformaticsweb.co.nr: Open Access Bioinformatics Education Resource. URL http://bioinformaticsweb.net/.

[6] BISTI - Biomedical Information Science and Technology Initiative. URL http://www.bisti.nih.gov/.

[7] Nair, A. S. (2007). Computational biology & bioinformatics: a gentle overview. Communications of Computer Society of India, 2, 1-13.

[8] Bioinformatics - Wikipedia, the free encyclopedia. URL http://en.wikipedia.org/wiki/Bioinformatics.

[9] Brown, F. K. (1998). Chemoinformatics: what is it and how does it impact drug discovery. Annual reports in medicinal chemistry, 33, 375-384.

[10] Paris G. (1999). Meeting of the American Chemical Society, citirao W.Warr. URL http://www.warr.com/warrzone2000.html.

[11] Gasteiger, J., & Funatsu, K. (2006). Chemoinformatics–An Important Scientific Discipline. Journal of Computer Chemistry, Japan, 5(2), 53-58.

[12] Drug design - Wikipedia, the free encyclopedia. URL http://en.wikipedia.org/wiki/Drug_design.

[13] Gasteiger, J., & Engel, T. (Eds.). (2003). Chemoinformatics: A Textbook. Wiley-VHC.

[14] Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., & Schacht, A. L. (2010). How to improve R&D productivity:

the pharmaceutical industry's grand challenge. Nature reviews Drug discovery, 9(3), 203-214.

[15] Prentis, R. A., Lis, Y. & Walker, S. R. (1988). Pharmaceutical innovation by the seven UK-owned pharmaceutical companies (1964-1985). British journal of clinical pharmacology, 25(3), 387-396.

[16] Hardman, J. G., Limbird, L. E., Molinoff, P. B., Ruddon, R. W., & Gilman, A. G. (1996). The pharmacologic basis of therapeutics. McGraw-Hill, Health Professions Division, 1465-1476.

[17] Sadowski, J., & Kubinyi, H. (1998). A scoring scheme for discriminating between drugs and nondrugs. Journal of medicinal chemistry, 41(18), 3325-3329.

[18] Walters, W. P., Murcko, A. A., & Murcko, M. A. (1999). Recognizing molecules with drug-like properties. Current opinion in chemical biology, 3(4), 384-387.

[19] Walters, W. P., Stahl, M. T., & Murcko, M. A. (1998). Virtual screening-an overview. Drug Discovery Today, 3(4), 160-178.

[20] Swofford, D., Olsen, G., Waddell, P., & Hillis, D.M. (1996) Phylogenetic inference. In Hillis, Moritz and Mable (eds), Molecular Systematics, 2nd edition. Sinauer, Sunderland, Massachusetts, pp. 407-511.

[21] Hershkovitz, M.A., & Leipe, D.D. (2006) Phylogenetic Analysis, in Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Volume 39 (eds A. D. Baxevanis and B. F. F. Ouellette), John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9780470110607.ch9.

[22] Jukes, T.H., & Cantor, C.R. (1969) Mammalian Protein Metabolism, Academic Press, New York, pp. 21-132.

[23] Kimura, M. (1980) A simple model for estimating evolutionary rates of base substitiutions through comparative studies of nucleotide sequences. J. Mol. Evol., 16, 111-120.

[24] Kishino, H., & Hasegawa, M. (1989) Evolution of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoida. J. Mol. Evol., 29, 170-179.

[25] Lake, J.A. (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl Acad. Sci. USA, 91, 1455-1459.

[26] Sokal, R., & Michener, C. (1958). A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin 38: 1409-1438.

[27] Sneath, P.H., & Sokal, R.R. (1973) Numerical Taxonomy, W. H. Freeman and Company, San Francisco, pp. 230-234.

[28] Saitou, N., & Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol., 4, 406-425.

[29] Fitch, W.M., & Margoliash, E. (1967) Construction of phylogenetic trees. Science 155: 279-284.

[30] Fitch, W.M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool., 35, 406-416.

[31] Sankoff, D., & Cedergren, R.J.,Time warps, string edits, and macromolecules: The theory and practice of sequence comparison (Addison-Wesley, London, 1983) pp. 253-264.

[32] Camin, J., & Sokal, R. (1965) A method for deducing branching sequences in phylogeny. Evolution, 19, 311-326.

[33] Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Systematic Biology, 22(3), 240-249.

[34] Larget, B., & Simon, D.L. (1999) Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. Mol Biol Evol 16(6): 750.

[35] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1), 97-109.

[36] David, W.M. (2001) Bioinformatics: sequence and genome analysis. Cold Spring Harbor Laboratory Press. pp. 6.

[37] Felsenstein, J. (1989) PHYLIP-Phylogeny Inference Package (Version 3.2). Cladistics, 5, 164-166.

[38] Thompson, J.D., Gibson, T., & Higgins, D.G. (2002) Multiple sequence alignment using ClustalW and ClustalX. Current protocols in bioinformatics, 2-3.

[39] Kumar, S., Nei, M., Dudley, J., & Tamura, K. (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings in bioinformatics, 9(4), 299-306.

[40] Swofford, D.L. (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

[41] Adachi, J., & Hasegawa, M. (1992) MOLPHY: Programs for molecular phylogenetics. I. PROTML: Maximum likelihood inference of protein phylogeny. In Computer Science Monographs, Vol.

[42] Schmidt, H.A., Strimmer, K., Vingron, M., & Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics, 18(3), 502-504.

[43] Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution, 24(8), 1586-1591.

[44] Ronquist, F., & Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 19(12), 1572-1574.

[45] Drummond, A. J., & Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology, 7(1), 214.

[46] Guindon, S., & Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic biology, 52(5), 696-704.

[47] Williams, T.L., & Moret, B.M. (2003) An investigation of phylogenetic likelihood methods. In Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on, pp. 79-86.

[48] Roch, S. (2010) Toward extracting all phylogenetic information from matrices of evolutionary distances. Science, 327(5971), 1376-1379.

[49] Bhardwaj G., , Ko, K.D., Hong, Y., Zhang, Z., Ho, N.L., Chintapalli, S.V.... & Rossum, D.B. (2012). PHYRN: A Robust Method for Phylogenetic Analysis of Highly Divergent Sequences. PloS one, 7(4), e34261.

[50] Brocchieri, L. (2001) Phylogenetic inferences from molecular sequences: review and critique. Theoretical population biology, 59(1), 27-40.

[51] Albayrak, A., Otu, H.H., & Sezerman, U.O. (2010) Clustering of protein families into functional subtypes using Relative Complexity Measure with reduced amino acid alphabets. BMC bioinformatics, 11(1), 428.

[52] Chen, X., & Tompa, M. (2010) Comparative assessment of methods for aligning multiple genome sequences. Nature biotechnology, 28(6), 567-572.

[53] Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research, 22(22), 4673-4680.

[54] Tommaso, P. D., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.M., … & Notredame, C. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic acids research, 39(suppl 2), W13-W17.

[55] Katoh, K., Kuma, K. I., Toh, H., & Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic acids research, 33(2), 511-518.

[56] Zhang, S., & Wang, T. (2010). A novel alignment-free method for phylogenetic analysis of protein sequences. In Proceedings of the 10th WSEAS international conference on Applied computer science. (pp. 67-71). World Scientific and Engineering Academy and Society (WSEAS).

[57] Zhang, S., & Wang, T. (2010). Phylogenetic analysis of protein sequences based on conditional LZ complexity. MATCH Commun. Math. Comput. Chem, 63(3), 701-716.

[58] Otu, H.H., & Sayood, K. (2003). A new sequence distance measure for phylogenetic tree construction. Bioinformatics, 19(16), 2122-2130.

[59] Carr, K., Murray, E., Armah, E., He, R.L., & Yau, S.S.T. (2010). A rapid method for characterization of protein relatedness using feature vectors. PloS one, 5(3), e9550.

[60] Bakis, Y., Otu, H.H., Tasçi, N., Meydan, C., Yüzbasioglu, S., & Sezerman, O.U. (2013). Testing robustness of relative complexity measure method constructing robust phylogenetic trees for Galanthus L. Using the relative complexity measure. BMC bioinformatics, 14(1), 20.

[61] Emil Fischer (2006). Einfluss der Configuration auf die Wirkung der Enzyme. Wiley-VCH, 27(3): 2985-2993. doi:10.1002/cber.18940270364.

[62] Worldwide Protein Data Bank. URL http://www.wwpdb.org/.

[63] RCSB PDB - Holdings Report. URL http://www.rcsb.org/pdb/statistics/holdings.do.

[64] Goodsell, D. S., Morris, G. M., & Olson, A. J. (1996). Automated docking of flexible ligands: applications of AutoDock. Journal of Molecular Recognition, 9(1), 1-5.

[65] Carlson, H. A., & McCammon, J. A. (2000). Accommodating protein flexibility in computational drug design. Molecular pharmacology, 57(2), 213-218.

[66] Stoddard, B. L., & Koshland, D. E. (1992). Prediction of the structure of a receptor–protein complex using a binary docking method.

[67] Halperin, I., Ma, B., Wolfson, H., & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins: Structure, Function, and Bioinformatics, 47(4), 409-443.

[68] Lin, S. L., Nussinov, R., Fischer, D., & Wolfson, H. J. (1994). Molecular surface representations by sparse critical points. Proteins: Structure, Function, and Bioinformatics, 18(1), 94-101.

[69] Lin, S. L., & Nussinov, R. (1996). Molecular recognition via face center representation of a molecular surface. Journal of Molecular Graphics, 14(2), 78-90.

[70] Connolly, M. L. (1983). Analytical molecular surface calculation. Journal of Applied Crystallography, 16(5), 548-558.

[71] Teschner, M., & Henn, C. (1995). Mapping volumetric properties on molecular surfaces in real-time. In System Sciences, 1995. Vol. V. Proceedings of the Twenty-Eighth Hawaii International Conference on (Vol. 5, pp. 265-272). IEEE.

[72] Srinark, T., & Kambhamettu, C. (2003, December). An approach for 3d segmentation on multiresolution surfaces. In Proceedings of International Conference on Intelligent Technologies (pp. 384-393).

[73] Taylor, R. D., Jewsbury, P. J., & Essex, J. W. (2002). A review of protein-small molecule docking methods. Journal of computer-aided molecular design, 16(3), 151-166.

[74] Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. Nature reviews Drug discovery, 3(11), 935-949.

[75] Scoring functions for docking - Wikipedia, the free encyclopedia. URL http: //en.wikipedia.org/wiki/Scoring_functions_for_docking.

[76] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., ... & Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society, 117(19), 5179-5197.

[77] Jorgensen, W. L., & Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. Journal of the American Chemical Society, 110(6), 1657-1666.

[78] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. Journal of computational chemistry, 4(2), 187-217.

[79] Nemethy, G., Gibson, K. D., Palmer, K. A., Yoon, C. N., Paterlini, G., Zagari, A., ... & Scheraga, H. A. (1992). Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. The Journal of Physical Chemistry, 96(15), 6472-6484.

[80] Oostenbrink, C., Villa, A., Mark, A. E., & Van Gunsteren, W. F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. Journal of computational chemistry, 25(13), 1656-1676.

[81] Rognan, D., Lauemøller, S. L., Holm, A., Buus, S., & Tschinke, V. (1999). Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. Journal of medicinal chemistry, 42(22), 4650-4658.

[82] Kramer, B., Rarey, M., & Lengauer, T. (1999). Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking. Proteins: Structure, Function, and Bioinformatics, 37(2), 228-241.

[83] Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., & Mee, R. P. (1997). Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. Journal of computer-aided molecular design, 11(5), 425-445.

[84] Böhm, H. J. (1994). The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. Journal of Computer-Aided Molecular Design, 8(3), 243-256.

[85] Rarey, M., Kramer, B., Lengauer, T., & Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. Journal of molecular biology, 261(3), 470-489.

[86] Gohlke, H., Hendlich, M., & Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. Journal of molecular biology, 295(2), 337-356.

[87] DeWitte, R. S., & Shakhnovich, E. I. (1996). SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. Journal of the American Chemical Society, 118(47), 11733-11744.

[88] Muegge, I., & Martin, Y. C. (1999). A general and fast scoring function for protein-ligand interactions: a simplified potential approach. Journal of medicinal chemistry, 42(5), 791-804.

[89] Wang, R., Lai, L., & Wang, S. (2002). Further development and validation of empirical scoring functions for structure-based binding affinity prediction. Journal of Computer-Aided Molecular Design, 16(1), 11-26.

[90] Shoichet, B. K., Kuntz, I. D., & Bodian, D. L. (1992). Molecular docking using shape descriptors. Journal of Computational Chemistry, 13(3), 380-397.

[91] Ewing, T. J., & Kuntz, I. D. (1997). Critical evaluation of search algorithms for automated molecular docking and database screening. Journal of Computational Chemistry, 18(9), 1175-1189.

[92] Ewing, T. J., Makino, S., Skillman, A. G., & Kuntz, I. D. (2001). DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. Journal of computer-aided molecular design, 15(5), 411-428.

[93] Schnecke, V., & Kuhn, L. A. (2000). Virtual screening with solvation and ligand-induced complementarity. Perspectives in drug discovery and design, 20(1), 171-190.

[94] Fast Rigid Exhaustive Docking. 2008. OpenEye Scientific Software, Santa Fe, New Mexico.

[95] Jones, G., Willett, P., Glen, R. C., Leach, A. R., & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. Journal of molecular biology, 267(3), 727-748.

[96] Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., ... & Shenkin, P. S. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. Journal of medicinal chemistry, 47(7), 1739-1749.

[97] Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. Journal of computational chemistry, 30(16), 2785-2791.

[98] Venkatachalam, C. M., Jiang, X., Oldfield, T., & Waldman, M. (2003). LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. Journal of Molecular Graphics and Modelling, 21(4), 289-307.

[99] Abagyan, R., Totrov, M., & Kuznetsov, D. (1994). ICM—a new method for protein modeling and design: applications to docking and structure prediction

from the distorted native conformation. Journal of computational chemistry, 15(5), 488-506.

[100] McMartin, C., & Bohacek, R. S. (1997). QXP: powerful, rapid computer algorithms for structure-based drug design. Journal of computer-aided molecular design, 11(4), 333-344.

[101] Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham III, T. E., DeBolt, S., ... & Kollman, P. (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Computer Physics Communications, 91(1), 1-41.

[102] Liu, M., & Wang, S. (1999). MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. Journal of computer-aided molecular design, 13(5), 435-451.

[103] Trosset, J. Y., & Scheraga, H. A. (1999). PRODOCK: software package for protein modeling and docking. Journal of computational chemistry, 20(4), 412-427.

[104] Clark, K. P. (1995). Flexible ligand docking without parameter adjustment across four ligand–receptor complexes. Journal of Computational Chemistry, 16(10), 1210-1226.

[105] Taylor, J. S., & Burnett, R. M. (2000). DARWIN: a program for docking flexible molecules. Proteins: Structure, Function, and Bioinformatics, 41(2), 173-191.

[106] Mizutani, M. Y., Tomioka, N., & Itai, A. (1994). Rational automatic search method for stable docking models of protein and ligand. Journal of molecular biology, 243(2), 310-326.

[107] Welch, W., Ruppert, J., & Jain, A. N. (1996). Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. Chemistry & biology, 3(6), 449-462.

[108] Miller, M. D., Kearsley, S. K., Underwood, D. J., & Sheridan, R. P. (1994). FLOG: a system to select ‘quasi-flexible’ligands complementary to a receptor of

known three-dimensional structure. Journal of computer-aided molecular design, 8(2), 153-174.

[109] Pang, Y. P., Perola, E., Xu, K., & Prendergast, F. G. (2001). EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. Journal of computational chemistry, 22(15), 1750-1771.

[110] Burkhard, P., Taylor, P., & Walkinshaw, M. D. (1998). An example of a protein ligand found by database mining: description of the docking method and its verification by a 2.3 Å X-ray structure of a Thrombin-Ligand complex. Journal of molecular biology, 277(2), 449-466.

[111] Gabb, H. A., Jackson, R. M., & Sternberg, M. J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. Journal of molecular biology, 272(1), 106-120.

[112] Metaphorics LLC, Mission Viejo, CA 92691.

[113] Baxter, C. A., Murray, C. W., Clark, D. E., Westhead, D. R., & Eldridge, M. D. (1998). Flexible docking using tabu search and an empirical estimate of binding affinity. Proteins: Structure, Function, and Bioinformatics, 33(3), 367-382.

[114] Accelrys Inc., San Diego, CA 92121.

[115] Luty, B. A., Wasserman, Z. R., Stouten, P. F., Hodge, C. N., Zacharias, M., & McCammon, J. A. (1995). A molecular mechanics/grid method for evaluation of ligand–receptor interactions. Journal of Computational Chemistry, 16(4), 454-464.

[116] Smoluchowski, M. V. (1917). Versuch einer mathematischen Theorie der Koagulationskinetik kolloider Lösungen. Z. phys. Chem, 92(129-168), 9.

[117] Northrup, S. H., & Erickson, H. P. (1992). Kinetics of protein-protein association explained by Brownian dynamics computer simulation. Proceedings of the National Academy of Sciences, 89(8), 3338-3342.

[118] Sharp, K., Fine, R., & Honig, B. (1987). Computer simulations of the diffusion of a substrate to an active site of an enzyme. Science, 236(4807), 1460-1463.

[119] Wiegel, F. W., & DeLisi, C. (1982). Evaluation of reaction rate enhancement by reduction in dimensionality. American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, 243(5), R475-R479.

[120] McCloskey, M. A., & Poo, M. M. (1986). Rates of membrane-associated reactions: reduction of dimensionality revisited. The Journal of cell biology, 102(1), 88-96.

[121] Peters, R. (2005). Translocation Through the Nuclear Pore Complex: Selectivity and Speed by Reduction-of-Dimensionality. Traffic, 6(5), 421-427.

[122] Brune, D., & Kim, S. (1994). Hydrodynamic steering effects in protein association. Proceedings of the National Academy of Sciences, 91(8), 2930-2934.

[123] Fröhlich, H. (1968). Long-range coherence and energy storage in biological systems. International Journal of Quantum Chemistry, 2(5), 641-649.

[124] Fröhlich, H. (1970). Long Range Coherence and the Action of Enzymes. Nature, 228, 1093.

[125] Fröhlich, H. (1975). The extraordinary dielectric properties of biological materials and the action of enzymes. Proceedings of the National Academy of Sciences, 72(11), 4211-4215.

[126] Little, W. A. (1964). Possibility of synthesizing an organic superconductor. Physical Review, 134(6A), A1416.

[127] Veljkovic, V. (1980) Theoretical approach to preselection of cancerogens and chemical carcinogenesis. Gordon & Breach, New York

[128] Veljkovic, V. J., & Lalovic, D. I. (1976). Theoretical prediction of mutagenicity and carcinogenicity of chemical substances. Cancer biochemistry biophysics, 1(6), 295.

[129] Veljkovic, V., & Slavic, I. (1972). Simple General-Model Pseudopotential. Physical Review Letters, 29, 105-107.

[130] Veljkovic, V. (1973). The dependence of the Fermi energy on the atomic number. Physics Letters A, 45(1), 41-42.

[131] Veljkovic, V., & Lalovic, D. I. (1973). General model pseudopotential for positive ions. Physics Letters A, 45(1), 59-60.

[132] Veljkovic, V., & Lalovic, D. I. (1978). Correlation between the carcinogenicity of organic substances and their spectral characteristics. Experientia, 34(10), 1342-1343.

[133] Ajdacic, V., & Veljkovic, V. (1978). Antibiotic activity of organic compounds and their average quasi-valence number. Experientia, 34(5), 633-635.

[134] Veljkovic, V., & Ajdacic, V. (1978). Cytostatic activity of organic compounds and their average quasi-valence number. Experientia, 34(5), 639-641.

[135] Veljkovic, V., Mouscadet, J. F., Veljkovic, N., Glisic, S., & Debyser, Z. (2007). Simple criterion for selection of flavonoid compounds with anti-HIV activity. Bioorganic & medicinal chemistry letters, 17(5), 1226-1232.

[136] Tintori, C., Manetti, F., Veljkovic, N., Perovic, V., Vercammen, J., Hayes, S., ... & Botta, M. (2007). Novel virtual screening protocol based on the combined use of molecular modeling and electron-ion interaction potential techniques to design HIV-1 integrase inhibitors. Journal of chemical information and modeling, 47(4), 1536-1544.

[137] GenBank Home. URL http://www.ncbi.nlm.nih.gov/genbank.

[138] NCBI - National Center for Biotechnology Information. URL http://www.ncbi. nlm.nih.gov.

[139] EMBL - European Molecular Biology Laboratory. URL http://www.embl.org.

[140] DDBJ - DNA Data Bank of Japan. URL http://www.ddbj.nig.ac.jp.

[141] INSDC - International Nucleotide Sequence Database Collaboration. URLs http://insdc.org ; http://www.ncbi.nlm.nih.gov/collab.

[142] Cochrane, G., Karsch-Mizrachi, I., & Nakamura, Y. (2011). The international nucleotide sequence database collaboration. Nucleic acids research, 39(suppl 1), D15-D18.

[143] NBRF - National Biomedical Research Foundation. URL http://pir.georgetown. edu.

[144] Dayhoff, M. O. (Ed.). (1973). Atlas of protein sequence and structure (Vol. 5). National Biomedical Research Foundation.

[145] SIB - Swiss Institute of Bioinformatics. URL http://www.isb-sib.ch.

[146] EBI - European Bioinformatics Institute. URL http://www.ebi.ac.uk.

[147] NIH - National Institute of Health. URL http://www.nih.gov.

[148] UNIPROT. URL http://www.uniprot.org.

[149] GenomeNet. Kyoto University Bioinformatics Center. URL http://www.genome.jp/en/about.html.

[150] GQuery: Global Cross-database NCBI search. URL http://www.ncbi.nlm.nih.gov/sites/gquery.

[151] DBGET - GenomeNet. URL http://www.genome.jp/en/gn_dbget.html.

[152] IGSP - Influenza Genome Sequencing Project. URL http://www.niaid.nih.gov/labsandresources/resources/dmid/gsc/influenza.

[153] NIAID - National Institute of Allergy and Infectious Diseases. URL http://www.niaid.nih.gov.

[154] IRD - Influenza Research Database. URL http://www.fludb.org.

[155] Influenza Virus Resource. URL http://www.ncbi.nlm.nih.gov/genomes/FLU.

[156] ChemDB. Division of AIDS Anti-HIV/OI/TB Therapeutics Database. URL http://chemdb.niaid.nih.gov.

[157] PubChem Project. URL http://pubchem.ncbi.nlm.nih.gov.

[158] ASINEX. URL http://www.asinex.com.

[159] Asinex Download Zone. URL http://www.asinex.com/download-zone.html.

[160] Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced drug delivery reviews, 23(1), 3-25.

[161] Dalby, A., Nourse, J. G., Hounshell, W. D., Gushurst, A. K., Grier, D. L., Leland, B. A., & Laufer, J. (1992). Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. Journal of chemical information and computer sciences, 32(3), 244-255.

[162] Daylight Theory: SMILES. URL http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html.

[163] Anderson, E., Veith, G. D., & Weininger, D. (1987). SMILES, a Line Notation and Computerized Interpreter for Chemical Structures. US Environmental Protection Agency, Environmental Research Laboratory.

[164] Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences, 28(1), 31-36.

[165] Weininger, D., Weininger, A., & Weininger, J. L. (1989). SMILES. 2. Algorithm for generation of unique SMILES notation. Journal of Chemical Information and Computer Sciences, 29(2), 97-101.

[166] Milic, L., Dobrosavljevic, Z., & Paunovic, Đ. (1999). Uvod u digitalnu obradu signala. Elektrotehnicki fakultet.

[167] Duhamel, P., & Hollmann, H. (1984). Split-radix FFT algorithm. Electron. Lett. 20 (1), 14-16.

[168] Ifeachor, E. C., & Jervis, B. W. (2002). Digital signal processing: a practical approach. Pearson Education.

[169] Lazovic, J. (1996). Selection of amino acid parameters for Fourier transform-based analysis of proteins. Computer applications in the biosciences: CABIOS, 12(6), 553-562.

[170] Veljkovic, V., Cosic, I., & Lalovic, D. (1985). Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?. Biomedical Engineering, IEEE Transactions on, (5), 337-341.

[171] Cosic, I., Nešic, D., Pavlovic, M., & Williams, R. (1986). Enhancer binding proteins predicted by informational spectrum method. Biochemical and biophysical research communications, 141(2), 831-838.

[172] Veljkovic V, Metlas R (1988) Identification of nanopeptide from HTLV3., LAV and ARV-2 envelope gp120 determining binding to T4 cell surface protein. Cancer Biochem Biophys, 10:91–106.

[173] Skerl, V., & Pavlovic, M. (1988). Thymopoietins and long postsynaptic neurotoxins share common information in their primary structures. FEBS letters, 239(1), 141-146.

[174] Cosic, I., Pavlovic, M., & Vojisavljevic, V. (1989). Prediction of 'hot spots' in interleukin-2 based on informational spectrum characteristics of growth-regulating factors. Comparison with experimental data. Biochimie, 71(3), 333-342.

[175] Lalovic, D., & Veljkovic, V. (1990). The global average DNA base composition of coding regions may be determined by the electron-ion interaction potential. Biosystems, 23(4), 311-316.

[176] Cosic, I., & Hearn, M. (1991) J. Mol. Recognit., 4, 57-62.Cosic, I., & Hearn, M. T. (1991). 'Hot spot' amino acid distribution in Ha-ras oncogene product p21: Relationship to guanine binding site. Journal of Molecular Recognition, 4(2-3), 57-62.

[177] Cosic, I., Hodder, A. N., Aguilar, M. I., & Hearn, M. T. (1991). Resonant recognition model and protein topography. Model studies with myoglobin, hemoglobin and lysozyme. European journal of biochemistry/FEBS, 198(1), 113-119.

[178] Veljkovic, V., Metlas, R., Raspopovic, J., & Pongor, S. (1992). Spectral and sequence similarity between vasoactive intestinal peptide and the second conserved region of human immunodeficiency virus type 1 envelope glycoprotein (gp120): possible consequences on prevention and therapy of AIDS. Biochemical and biophysical research communications, 189(2), 705-710.

[179] Cosic, I. (1994). Macromolecular bioactivity: is it resonant interaction between macromolecules - theory and applications. Biomedical Engineering, IEEE Transactions on, 41(12), 1101-1114.

[180] Cosic, I., Drummond, A. E., Underwood, J. R., & Hearn, M. T. (1994). In vitro inhibition of the actions of basic FGF by a novel 16 amino acid peptide. Molecular and cellular biochemistry, 130(1), 1-9.

[181] Cosic, I., & Birch, S. (1994, November). Photoreceptors having similar structure but different absorptions can be distinguished using the resonant recognition

model. In Engineering in Medicine and Biology Society, 1994. Engineering Advances: New Opportunities for Biomedical Engineers. Proceedings of the 16th Annual International Conference of the IEEE (pp. 265-266). IEEE.

[182] Nair, T. M., Tambe, S. S., & Kulkarni, B. D. (1994). Application of artificial neural networks for prokaryotic transcription terminator prediction. FEBS letters, 346(2), 273-277.

[183] Veljkovic, V., Johnson, E., & Metlas, R. (1995). Analogy of HIV-1 to oncogenic viruses: possible implications for the pathogenesis of AIDS. The Cancer journal, 8(6), 308-314.

[184] Cosic, I. (1995). Virtual spectroscopy for fun and profit. Nature BioTechnology, 13(3), 236-238.

[185] Birch, S., West, R., & Cosic, I. (1995). Preliminary expansion of the resonant recognition model to incorporate multi variable analysis. Australasian physical & engineering sciences in medicine/supported by the Australasian College of Physical Scientists in Medicine and the Australasian Association of Physical Sciences in Medicine, 18(4), 197.

[186] Krsmanovic, V., Biquard, J. M., Sikorska-Walker, M., Cosic, I., Desgranges, C., Trabaud, M. A., ... & Hearn, M. T. W. (1998). Investigations into the cross-reactivity of rabbit antibodies raised against nonhomologous pairs of synthetic peptides derived from HIV-1 gp120 proteins. The Journal of peptide research, 52(5), 410-420.

[187] Fang, Q., & Cosic, I. (1998). Protein structure analysis using the resonant recognition model and wavelet transforms. Australasian physical & engineering sciences in medicine/supported by the Australasian College of Physical Scientists in Medicine and the Australasian Association of Physical Sciences in Medicine, 21(4), 179.

[188] Hejase de Trad, C., Fang, Q., & Cosic, I. (2000). The resonant recognition model (RRM) predicts amino acid residues in highly conserved regions of the hormone prolactin (PRL). Biophysical Chemistry, 84(2), 149-157.

[189] Hejase de Trad, C., Fang, Q., & Cosic, I. (2002). Protein sequence comparison based on the wavelet transform approach. Protein Engineering, 15(3), 193-203.

[190] Parbhane, R. V., Unniraman, S., Tambe, S. S., Nagaraja, V., & Kulkarni, B. D. (2000). Optimum DNA curvature using a hybrid approach involving an artificial neural network and genetic algorithm. Journal of Biomolecular Structure and Dynamics, 17(4), 665-672.

[191] Murakami, M. (2000). Resonant Recognition Model of Neuropeptide Y Family: Hot Spot Amino Acid Distribution in the Sequences. Journal of Protein Chemistry, 19(7), 609-613.

[192] Durbin, R. (Ed.). (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge university press.

[193] Minkowski, H. (1953). Geometrie der Zahlen. Chelsea.

[194] Schwarz, R., & Dayhoff, M. (1979). Matrices for detecting distant relationships. In Dayhoff M., editor, Atlas of protein sequences, pages 353-58. National Biomedical Research Foundation.

[195] Jones, D.T., Taylor, W.R., & Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biosciences 8: 275-282.

[196] Murtagh, F. (1984). Complexities of Hierarchic Clustering Algorithms: the state of the art. Computational Statistics Quarterly 1: 101-113.

[197] Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol 28:2731-9.

[198] Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32(5), 1792-1797.

[199] Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J.,... & Springer, M. S. (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science, 294(5550), 2348-2351.

[200] Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., O'Brien, & S. J. (2001) Molecular phylogenetics and the origins of placental mammals. Nature, 409(6820), 614-618.

[201] V. Ranwez, F. Delsuc, S. Ranwez, K. Belkhir, M. K. Tilak, E. J. Douzery (2007) OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. BMC Evolutionary Biology, 7(1), 241.

[202] Olivier, M., Eeles, R., Hollstein, M., Khan, M. A., Harris, C. C., & Hainaut, P. (2002). The IARC TP53 database: new online mutation analysis and recommendations to users. Human mutation, 19(6), 607-614.

[203] Veljkovic, N., & Perovic, V. (2006). In silico criterion for prediction of effects of p53 gene missense mutations on p53-Mdm2 feedback loop. Protein and Peptide Letters, 13(8), 807-814.

[204] Wolfram Mathematica 9 Documentation. Distance and Similarity Measures. URL http://reference.wolfram.com/mathematica/guide/DistanceAndSimilarity Measures.html.

[205] Krause, E. (1987). Taxicab geometry: An adventure in non-Euclidean geometry. DoverPublications. com.

[206] Black, P. E. (2004). „Manhattan distance“ in Dictionary of algorithms and data structures. National Institute of Standards and Technology.

[207] Deza, M. M., & Deza, E. (2009). Encyclopedia of distances (pp. 1-583) page 94. Springer Berlin Heidelberg.

[208] Rudin, W. (1986). Real and complex analysis (3rd). New York: McGraw-Hill Inc.

[209] Androutsos, D., Plataniotis, K. N., & Venetsanopoulos, A. N. (1998, December). Vector angular distance measure for indexing and retrieval of color. In Electronic Imaging '99 (pp. 604-613). International Society for Optics and Photonics.

[210] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.

[211] Fulekar, M. H. (Ed.). (2009). Bioinformatics: Applications in life and environmental sciences. page 110. Springer.

[212] Jurman, G., Riccadonna, S., Visintainer, R., & Furlanello, C. (2009). Canberra distance on ranked lists. In Proceedings, Advances in Ranking–NIPS 09 Workshop (pp. 22-27).

[213] Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. Ecological monographs, 27(4), 325-349.

[214] Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. Software: Practice and experience, 21(11), 1129-1164.

[215] Walshaw, C. (2001, January). A multilevel algorithm for force-directed graph drawing. In Graph Drawing (pp. 171-182). Springer Berlin Heidelberg.

[216] Hu, Y. (2005). Efficient, high-quality force-directed graph drawing. Mathematica Journal, 10(1), 37-71.

[217] Tutte, W. T. (1963). How to draw a graph. Proc. London Math. Soc, 13(3), 743-768.

[218] P. Eades (1984). A heuristic for graph drawing. Congressus Numerantium, 42:149–160.

[219] Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. Information processing letters, 31(1), 7-15.

[220] S. G. Kobourov, "Force-Directed Drawing Algorithms," In Tamassia, R. (Ed.). (2007). Handbook of graph drawing and visualization. p. 383-408, Chapman & Hall/CRC.

[221] Kobourov, S. G. (2012). Spring Embedders and Force Directed Graph Drawing Algorithms. arXiv preprint arXiv:1201.3011.

[222] Gemovic, B., Perovic, V., Glisic, S., Veljkovic, N. (2013). Feature-Based Classification of Amino Acid Substitutions outside Conserved Functional Protein Domains. Scientific World Journal, doi:10.1155/2013/948617.

[223] Glisic, S., Veljkovic, N., Cupic, S. J., Vasiljevic, N., Prljic, J., Gemovic, B., Perovic, V., Veljkovic, V. (2012). Assessment of Hepatitis C Virus Protein Sequences with Regard to Interferon/Ribavirin Combination Therapy Response in Patients with HCV Genotype 1b. Protein J, 31(2): 129-136.

[224] Veljkovic, V., Niman, H. L., Glisic, S., Veljkovic, N., Perovic, V., & Muller, C. P. (2009). Identification of hemagglutinin structural domain and polymorphisms which may modulate swine H1N1 interactions with human receptor. BMC Structural Biology, 9(1), 62.

[225] Veljkovic, N., Glisic, S., Prljic, J., Perovic, V., Botta, M., & Veljkovic, V. (2008). Discovery of new therapeutic targets by the informational spectrum method. Current Protein and Peptide Science, 9(5), 493-506.

[226] Glisic, S., Arrigo, P., Alavantic, D., Perovic, V., Prljic, J., & Veljkovic, N. (2008). Lipoprotein lipase: A bioinformatics criterion for assessment of mutations as a risk factor for cardiovascular disease. Proteins: Structure, Function, and Bioinformatics, 70(3), 855-862.

[227] Veljkovic, V., Veljkovic, N., Muller, C. P., Müller, S., Glisic, S., Perovic, V., & Köhler, H. (2009). Characterization of conserved properties of hemagglutinin of H5N1 and human influenza viruses: possible consequences for therapy and infection control. BMC Structural Biology, 9(1), 21.

[228] Rogers, G. N., & D'Souza, B. L. (1989). Receptor binding properties of human and animal H1 influenza virus isolates. Virology, 173(1), 317-322.

[229] Doliana, R., Veljkovic, V., Prljic, J., Veljkovic, N., De Lorenzo, E., Mongiat, M., ... & Colombatti, A. (2008). EMILINs interact with anthrax protective antigen and inhibit toxin action in vitro. Matrix Biology, 27(2), 96-106.

[230] Matrosovich, M., Tuzikov, A., Bovin, N., Gambaryan, A., Klimov, A., Castrucci, M. R., ... & Kawaoka, Y. (2000). Early alterations of the receptor-binding properties of H1, H2, and H3 avian influenza virus hemagglutinins after their introduction into mammals. Journal of virology, 74(18), 8502-8512.

[231] Matrosovich, M., Zhou, N., Kawaoka, Y., & Webster, R. (1999). The surface glycoproteins of H5 influenza viruses isolated from humans, chickens, and wild aquatic birds have distinguishable properties. Journal of virology, 73(2), 1146-1155.

[232] Veljkovic V, Veljkovic N and Metlas R (2004) Molecular makeup of HIV-1 envelope protein. Int Rev Immunol, 23:383–411.

[233] Veljkovic V, Metlas R (1990) Sequence similarity between HIV-1 envelope protein gp120 and human proteins, a new hypothesis on protective antibody production. Immunol Lett, 26:193–195.

[234] Veljkovic V, Metlas R, Kohler H, Urnovitz H, Prljic J, Veljkovic N, Johnson E and Muller S (2001) AIDS epidemic at the beginning of the third millennium, time for a new AIDS vaccine strategy. Vaccine, 19:1855–1562.

[235] Veljkovic V, Veljkovic N, Glisic S and Ho MW (2008) AIDS vaccine: efficacy, safety and ethics. Vaccine, 26:3072–3077.

[236] Köhler H, Müller S, Nara PL (1994) Deceptive imprinting in the immune response against HIV-1. Immunol Today, 15:475–478.

[237] Su, Y., Yang, H. Y., Zhang, B. J., Jia, H. L., & Tien, P. (2008). Analysis of a point mutation in H5N1 avian influenza virus hemagglutinin in relation to virus entry into live mammalian cells. Archives of virology, 153(12), 2253-2261.

[238] Goldberg, I. J., & Merkel, M. (2001). Lipoprotein lipase: physiology, biochemistry, and molecular biology. Frontiers in bioscience: a journal and virtual library, 6, D388.

[239] Hill, J. S., Yang, D., Nikazy, J., Curtiss, L. K., Sparrow, J. T., & Wong, H. (1998). Subdomain chimeras of hepatic lipase and lipoprotein lipase Localization of heparin and cofactor binding. Journal of Biological Chemistry, 273(47), 30979-30984.

[240] Shen, Y., Lookene, A., Nilsson, S., & Olivecrona, G. (2002). Functional analyses of human apolipoprotein CII by site-directed mutagenesis identification of residues important for activation of lipoprotein lipase. Journal of Biological Chemistry, 277(6), 4334-4342.

[241] Balasubramaniam A, Rechtin A, Mc Lean LR, Jackson RL, Demel RA (1986). Activation of lipoprotein lipase by N-alpha-palmitoyl (56-79) fragment of apolipoprotein C-II. Biochem Biophys Res Commun.137(3):1041-8

[242] Razzaghi, H., Day, B. W., McClure, R. J., & Kamboh, M. I. (2001). Structure–function analysis of D9N and N291S mutations in human lipoprotein lipase using molecular modelling. Journal of Molecular Graphics and Modelling, 19(6), 487-494.

[243] Kastan, M. B., Onyekwere, O., Sidransky, D., Vogelstein, B., & Craig, R. W. (1991). Participation of p53 protein in the cellular response to DNA damage. Cancer research, 51(23 Part 1), 6304-6311.

[244] Yonish-Rouach, E., Resnitzky, D., Lotem, J., Sachs, L., Kimchi, A., & Oren, M. (1991). Wild-type p53 induces apoptosis of myeloid leukaemic cells that is inhibited by interleukin-6. Nature, 352(6333), 345.

[245] Momand, J., Wu, H. H., & Dasgupta, G. (2000). MDM2-master regulator of the p53 tumor suppressor protein. Gene, 242(1), 15-29.

[246] Thomas, M., Kalita, A., Labrecque, S., Pim, D., Banks, L., & Matlashewski, G. (1999). Two polymorphic variants of wild-type p53 differ biochemically and biologically. Molecular and cellular biology, 19(2), 1092-1100.

[247] Dumont, P., Leu, J. J., Della Pietra, A. C., George, D. L., & Murphy, M. (2003). The codon 72 polymorphic variants of p53 have markedly different apoptotic potential. Nature genetics, 33(3), 357-365.

[248] Bougeard, G., Baert-Desurmont, S., Tournier, I., Vasseur, S., Martin, C., Brugieres, L., ... & Frebourg, T. (2006). Impact of the MDM2 SNP309 and p53 Arg72Pro polymorphism on age of tumour onset in Li-Fraumeni syndrome. Journal of medical genetics, 43(6), 531-533.

[249] Li, X., Dumont, P., Della Pietra, A., Shetler, C., & Murphy, M. E. (2005). The codon 47 polymorphism in p53 is functionally significant. Journal of Biological Chemistry, 280(25), 24245-24251.

[250] Marine, J. C., & Jochemsen, A. G. (2005). Mdmx as an essential regulator of p53 activity. Biochemical and biophysical research communications, 331(3), 750-760.

[251] Nikolaev, A. Y., Li, M., Puskas, N., Qin, J., & Gu, W. (2003). Parc: a cytoplasmic anchor for p53. Cell, 112(1), 29-40.

[252] Sengupta, S., & Wasylyk, B. (2001). Ligand-dependent interaction of the glucocorticoid receptor with p53 enhances their degradation by Hdm2. Genes & development, 15(18), 2367-2380.

[253] Vogelstein, B., & Kinzler, K. W. (1992). p53 function and dysfunction. Cell, 70(4), 523-526.

[254] Goh, H. S., Yao, J., & Smith, D. R. (1995). p53 point mutation and survival in colorectal cancer patients. Cancer research, 55(22), 5217-5221.

[255] Cosic, I., & Nesic, D. (1987). Prediction of 'hot spots' in SV40 enhancer and relation with experimental data. European Journal of Biochemistry, 170(1-2), 247-252.

[256] Venot, C., Maratrat, M., Sierra, V., Conseiller, E., & Debussche, L. (1999). Definition of a p53 transactivation function-deficient mutant and characterization of two independent p53 transactivation subdomains. Oncogene, 18(14), 2405-2410.

[257] Morris, S. M. (2002). A role for p53 in the frequency and mechanism of mutation. Mutation Research/Reviews in Mutation Research, 511(1), 45-62.

[258] Nowell, P. C. (1976). The clonal evolution of tumor cell populations. Science, 194(4260), 23-28.

[259] Jonason, A. S., Kunala, S., Price, G. J., Restifo, R. J., Spinelli, H. M., Persing, J. A., ... & Brash, D. E. (1996). Frequent clones of p53-mutated keratinocytes in normal human skin. Proceedings of the National Academy of Sciences, 93(24), 14025-14029.

[260] Wang TT, Parides MK, Palese P (2012) Seroevidence for H5N1 influenza infections in humans: meta-analysis. Science. 335(6075):1463.

[261] HAI - Influenza at the Human-Animal Interface. (World Health Organization). URL http://www.who.int/influenza/human_animal_interface/en/.

[262] Watanabe, Y., Ibrahim, M. S., Ellakany, H. F., Kawashita, N., Mizuike, R., Hiramatsu, H., Sriwilaijaroen, N., Takagi, T., Suzuki, Y., & Ikuta, K. (2011) Acquisition of human-type receptor binding specificity by new H5N1 influenza virus sublineages during their emergence in birds in Egypt. PLoS Pathog. 7(5):e1002068.

[263] Perovic, V. R. (2013). Novel algorithm for phylogenetic analysis of proteins: application to analysis of the evolution of H5N1 influenza viruses. Journal of Mathematical Chemistry, 1-18.

[264] Perovic, V. R., Muller, C. P., Niman, H. L., Veljkovic, N., Dietrich, U., Tosic, D. D., ... & Veljkovic, V. (2013). Novel Phylogenetic Algorithm to Monitor Human Tropism in Egyptian H5N1-HPAIV Reveals Evolution toward Efficient Human-to-Human Transmission. PloS one, 8(4), e61572.

[265] GISAID database. Global Initiative on Sharing All Influenza Data. URL http: //platform.gisaid.org.

[266] WHO. Cumulative number of confirmed human cases of avian influenza A(H5N1) reported to WHO. URL http://www.who.int/influenza/human_animal_ interface/H5N1_cumulative_table_archives/en/index.html.

[267] Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution, 39(4), 783-791.

[268] Seifert, M. H., Wolf, K., & Vitt, D. (2003). Virtual high-throughput *in silico* screening. Biosilico, 1(4), 143-149.

[269] Dixon, S., McDonald, S., & Roberts, J. (2002). The impact of HIV and AIDS on Africa's economic development. BMJ: British Medical Journal, 324(7331), 232.

[270] Hogg, R., Lima, V., Sterne, J. A., Grabar, S., Battegay, M., Bonarek, M., ... & May, M. (2008). Life expectancy of individuals on combination antiretroviral therapy in high-income countries: a collaborative analysis of 14 cohort studies. Lancet, 372(9635), 293-299.

[271] DiMasi, J. A. (2001). Risks in new drug development: approval success rates for investigational drugs. Clin.Pharma.Therapeutics. 69(5), 297-307.

[272] Maga, G., Veljkovic, N., Crespan, E., Spadari, S., Prljic, J., Perovic, V., ... & Veljkovic, V. (2013). New in silico and conventional in vitro approaches to advance HIV drug discovery and design. Expert opinion on drug discovery, 8(1), 83-92.

[273] Veljkovic, N., Glisic, S., Prljic, J., Perovic, V., & Veljkovic, V. (2012). Simple and General Criterion for" In Silico" Screening of Candidate HIV Drugs. Current pharmaceutical biotechnology, 14, 89.

[274] Mugnaini, C., Rajamaki, S., Tintori, C., Corelli, F., Massa, S., Witvrouw, M., ... & Botta, M. (2007). Toward novel HIV-1 integrase binding inhibitors: molecular modeling, synthesis, and biological studies. Bioorganic & medicinal chemistry letters, 17(19), 5370-5373.

[275] Veljkovic, V., Veljkovic, N., Este, J. A., Huther, A., & Dietrich, U. (2007). Application of the EIIP/ISM bioinformatics concept in development of new drugs. Current medicinal chemistry, 14(4), 441-453.

[276] Wang, Z., Bennett, E. M., Wilson, D. J., Salomon, C., & Vince, R. (2007). Rationally designed dual inhibitors of HIV reverse transcriptase and integrase. Journal of medicinal chemistry, 50(15), 3416-3419.

[277] Veljkovic, N., Glisic, S., Perovic, V., & Veljkovic, V. (2011). The role of long-range intermolecular interactions in discovery of new drugs. Expert Opinion on Drug Discovery, 6(12), 1263-1270.

[278] Kumarasamy, K. K., Toleman, M. A., Walsh, T. R., Bagaria, J., Butt, F., Balakrishnan, R., ... & Woodford, N. (2010). Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study. The Lancet infectious diseases, 10(9), 597-602.

Stručna biografija

Vladimir Perović je rođen 09.02.1976. godine u Beogradu. Završio je osnovnu školu „Drinka Pavlović“ kao đak generacije, sa osvojenom trećom nagradom na republičkom takmičenju iz matematike u osmom razredu. Matematičku gimnaziju u Beogradu završio je 1995. godine.

Školske 1995/1996 godine upisao je Matematički fakultet Univerziteta u Beogradu, smer računarstvo i informatika, na kome je diplomirao 19.10.2001. sa prosečnom ocenom 9,34/10,00 i stekao naziv Diplomirani matematičar.

Školske 2001/2002 upisao je magistarske studije na Matematičkom fakultetu Univerziteta u Beogradu, smer računarstvo i informatika, a 2009. godine se prebacuje na doktorske studije istog fakulteta, smer informatika.

Od 2002. godine zaposlen je u Laboratoriji za multidisciplinarna istraživanja i inženjering Instituta za nuklearne nauke „Vinča“. U zvanje istraživač-saradnik izabran je 2009. godine.

Do sada je publikovao 14 naučnih radova iz uže naučne oblasti, od toga: 7 radova u vrhunskim međunarodnim časopisima, 5 radova u istaknutim međunarodnim časopisima i 2 rada u međunarodnim časopisima. Pored toga, objavio je jedno tehničko rešenje - softver (algoritam), jedno saopštenje sa međunarodnih skupova štampano u celini i 6 saopštenje sa međunarodnih skupova štampano u izvodu.

Trenutno je angažovan na projektu „Primena EIIP/ISM bioinformatičke platforme u otkrivanju novih terapeutskih targeta i potencijalnih terapeutskih molekula“ (br. 173001) Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije.

Прилог 1.

Изјава о ауторству

Потписани-а Владимир Перовић

број индекса 2022 / 2009

Изјављујем

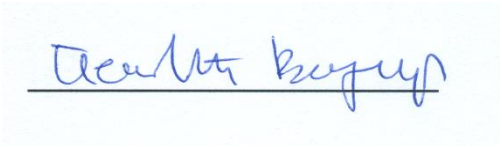
да је докторска дисертација под насловом

Развој мултифункционалне биоинформатичке платформе засноване на потенцијалу електрон-јон интеракције биолошких молекула

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, _____



Прилог 2.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора _____Владимир Перовић_____

Број индекса _____2022 / 2009_____

Студијски програм _____

Наслов рада __Развој мултифункционалне биоинформатичке платформе засноване на потенцијалу електрон-јон интеракције биолошких молекула_

Ментор __др Душан Тошић, редовни професор Математичког факултета Универзитета у Београду_____

Потписани/а _____Владимир Перовић_____

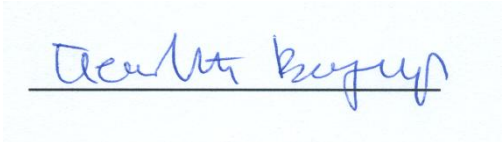
Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, _____



Прилог 3.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Развој мултифункционалне биоинформатичке платформе засноване на потенцијалу електрон-јон интеракције биолошких молекула

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство

2. Ауторство - некомерцијално

3. Ауторство – некомерцијално – без прераде

4. Ауторство – некомерцијално – делити под истим условима

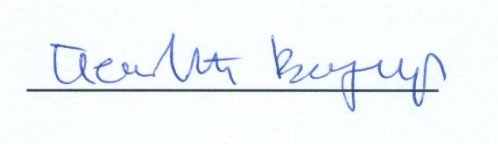
5. Ауторство – без прераде

6. Ауторство – делити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

Потпис докторанда

У Београду, _____



1. Ауторство - Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. Ауторство – некомерцијално. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. Ауторство - некомерцијално – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. Ауторство - некомерцијално – делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. Ауторство – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. Ауторство - делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.