

# Bioinformatika – najmlađa znanost o dekodiranju najstarijeg jezika: jezika gena

*Davor Juretić*

## **Bioinformatika, moderna biologija i Internet**

Ako napravimo istraživanje o tome koja je znanost najmanje zastupljena u Hrvatskoj, koja je najpristupačnija svima preko Interneta i za koju u bogatim zemljama svijeta postoji najveća potreba u mladim znanstvenim radnicima, mislim da bi odgovor uvijek bio isti: bioinformatika. No što je to bioinformatika? Kao što to naslov sugerira ta mlada znanost nastaje u sretnom braku informatike i moderne biologije koja želi dekodirati gene svih živih bića. Svi danas znamo da se geni mogu naći u nasljednim makromolekulama kao što je DNK, a posao bioinformatičara u zajednici s molekularnim biolozima je da otkriju gdje se geni točno nalaze u DNK, kako upravljaju sintezom proteina, te kakva je struktura i djelovanje tih proteina. Oko 40 tisuća gena, koliko ih ljudi imaju u svakoj svojoj stanici, zauzima samo mali postotak ukupne dužine DNK molekula u stanici, pa ih nije ni malo jednostavno naći. Od tog broja postoji nekoliko gena koji su ključni za razliku između nas i naših najbližih srodnika majmuna, ali još ni izdaleka ne znamo koji su to geni. Jedino smo sigurni za već dobro ispitane gene da su vrlo slični kod ljudi, majmuna, štakora, miševa, pa čak i vinskih mušica, koje su se od nas u evoluciji odvojile pred možda čak 500 miliona godina.

Priroda se stalno igra s genima tako da ih mijenja, recimo kroz spontane mutacije, ali kada jednom otkrije nešto što dobro radi to ostaje u biti isto kroz vrlo dugo vrijeme dok se čak i broj kontinenata na Zemlji promijenio. Većina pokušaja promjena takvih gena, koji održavaju red u stanici, oštro se kažnjava smrću te stanice, što naravno smjesta izbacuje promjenjene gene iz daljnje evolucije. To je ono što je već Charles Darwin, prije 150 godina, objasnio kao proces prirodnog odabiranja, dok se još nije ništa znalo o pravoj prirodi nositelja nasljednih osobina.

Zašto je bioinformatika najpristupačnija znanost, bez obzira što je najmanje poznata u našoj zemlji? Bioinformatika se razvijala zajedno sa Internetom, i bez Interneta ne bi mogla ni postojati, bar ne u današnjem obliku. U samom početku razvoja bioinformatike istraživači su se držali principa da sve glavne baze podataka iz bioinformatike budu svima slobodno dostupne. Taj princip poštuje se i danas, što je sigurno jako ubrzalo razvoj bioinformatike. Bioinformatika se najčešće ne spušta na nivo atoma, kao što to radi biofizika, niti joj je cilj objasniti mehanizme enzimatski kataliziranih reakcija, kao što je to često cilj u biokemiji. Dok je broj različitih atoma i stanja u

kojima se mogu nalaziti vrlo velik, a također i broj makromolekula od interesa za biokemiju, dotle bioinformatika uglavnom barata samo sa 20 ili čak samo sa četiri različita simbola za male molekule kao što su aminokiseline ili nukleotidi. Naime, priroda već više od tri milijarde godina koristi taj mali broj malih molekula da pomoću njih zapiše sve svoje poruke kako život treba da živi.

Genetički kod, u kojem su zapisane te poruke, nije ništa drugo nego skup pravila po kojima se redoslijed purinskih (adenin: A i gvanin: G) i pirimidinskih (timin: T i citozin: C) baza iz DNK sekvencije (4 razna slova) prevodi u slijed aminokiselina u proteinu (20 raznih slova). Taj kod u obliku tablice pridružuje jednu aminokiselinu svakom tripletu baza. Sigurno postoje i još nedovoljno poznata pravila kako se linearni slijed aminokiselina spontano pretvara u funkcionalnu 3D strukturu proteina, pa ta pravila neki bioinformatičari nazivaju drugom polovicom genetičkog koda.

Bioinformatičari je cilj da otkrije tu prastaru genetsku mudrost, a da bi to postigla mora znati pročitati poruke koje je priroda zapisala pomoću tih malih molekula. Broj različitih simbola, koje možemo smatrati slovima u rečenici jest malen, ali su rečenice koje priroda koristi vrlo dugačke. Dakle posao bioinformatičara donekle nalikuje poslu lingvista koji treba proučavati nepoznati jezik od kojeg imamo zapisani tekst ali ne razumijemo značenje rečenica. Taj jezik je još živ, to je najstariji živi fosil, jer se genetički kod ustalio prije više od tri milijarde godina.

Kako spoznajemo da smo našli neki tekst (na Internetu) koji je priroda zapisala u svom prastarom jeziku? Tekst koji će nam reći da se radi o bioinformatičari pojavljuje se recimo kao: DSLWWGTITLTTIGYGDKTP. U ovom slučaju priroda je koristila jezik u kojem postoji 20 raznih slova. Ova je poruka vrlo kratka, ima ukupno 20 slova, ali od toga samo 10 raznih slova. No svako se slovo mora nalaziti u točno određenom mjestu ove sekvencije da bi poruka zadržala svoj pravi smisao. Ako jedno od slova G u ovoj poruci postane S, dijete koje je od svojih roditelja dobilo takvu poruku s greškom oboljet će od nasljedne gluhoće. Slova u toj poruci su jednoslovni kodovi za aminokiseline. Glicin je G, leucin je L i tako dalje. Ova poruka nije ništa drugo nego povezan niz aminokiselina koje sve zajedno prave ono što biokemičari nazivaju polipeptid ili proteinska sekvencija (protein je bjelančevina).

Prije dvije godine razni genomni projekti izbacivali su svake sekunde po jednu novu sekvenciju ovakve vrste, sada ih već izbacuju desetak, a sutra će ih izbacivati 100 ili 1000 u svakoj sekundi i po danu i po noći. Jasno da je to ogromni posao za bioinformatičare da nađu neki smisao u svim tim biološkim sekvencijama. Prvo treba naći gdje su geni u dugačkoj DNK molekuli. Zatim treba jezik gena, koji koristi samo četiri slova, prevesti u jezik proteina koji koristi 20 slova i konačno treba otkriti kako izgleda i što radi taj protein. Nije čudo da je bioinformatičar jedno od danas najviše traženih zanimanja u svijetu. Ako uspijete sami riješiti “G se pretvorilo u S” zadatak,

tako da nadete koji glicin iz poruke izaziva nasljednu gluhoću kada se pretvori u serin, koristeći ništa drugo nego Internet, znači da imate smisla za to zanimanje.

## **Istraživanja u bioinformatici**

Bioinformatika je još vrlo mlada znanost, ne starija od nekih 15 godina. Kada sam pred 15 godina počeo razvijati jedan algoritam za predikciju sekundarne strukture membranskih proteina, nisam ni znao da se upravo počinjem baviti bioinformatikom. U to doba radio sam na Nacionalnim institutima za zdravlje u SAD, a to je najveća znanstvena institucija na svijetu, pa bih sigurno doznao da sudjelujem u rađanju jedne nove znanosti, da su to i drugi znanstvenici oko mene znali. Ono što sam radio iz znatiželje na predviđanju strukture proteina, tada se prepoznavalo kao računska biologija (computational biology) ili kao teorijska molekularna biologija. No, to je tipičan opis posla i za današnje bioinformatičare, koji uglavnom razvijaju nove alate za analizu i usporedbu bioloških sekvencija. Taj posao danas je neusporedivo opsežniji i zahtjevniji, jer su u međuvremenu razni projekti genoma izbacili ogroman broj DNK i proteinskih sekvenciji. Tako golemu količinu informacija nije moguće ni pohraniti ni iskorištavati bez da se koriste posebni alati (algoritmi, programi) napravljeni baš za tu svrhu. Poplava novih bioloških podataka, koja se i stalno pojačava, dovela je do toga da svako istraživanje u biološkim znanostima (pa i u medicini) mora započeti primjenom računala i pretraživanjem odgovarajućih banka podataka. Mlađim znanstvenicima danas je teško i zamisliti kako su se njihovi stariji kolege mogli snalaziti u istraživanju, kada nisu postojala osobna računala s pristupom Internetu na svakom radnom stolu.

Bioinformatika je danas najotvorenija znanost na svijetu i po dostupnosti bioinformatičkih alata, a ne samo baza podataka. Dio te znanosti je naravno upotrebljen za stvaranje komercijalnih paketa programa, no najveći dio bioinformatičkih alata slobodno je dostupan svima preko Interneta. Bioinformatika nije samo skup alata za biologe. U bioinformatičkom istraživanju vrijede isti principi kao i u svakom drugom istraživanju. Rezultati moraju jasno odgovoriti na pitanje koje istraživač postavlja. Pitanje, hipoteza, metode, cilj i rezultati istraživanja moraju svi biti dobro definirani i reproducibilni. Istraživač mora dobro poznavati osnove biologije/biokemije problema na kojem radi da bi uopće mogao postaviti dobro istraživačko pitanje iz područja bioinformatike. Naravno, mora biti vješt i u informatici i u matematici, tako da biološki problem može pretvoriti u algoritam, a algoritam u program na nekom od računalskih jezika. Također mora biti svijestan GIGO problema ("garbage in – garbage out"), koji je u bioinformatici vrlo čest. Naime, bioinformatički alati, kao i druge računalske metode, uvijek produciraju neke rezultate (što nije slučaj u eksperimentima), samo što su ti rezultati često smeće, odnosno nije lako razlikovati značajne rezultate od smeća. Ulazni podaci također često sadrže greške. Recimo, frekvencija

grešaka u DNK po jednoj bazi je u rasponu od 1/10000 za velike projekte genoma do 1/10 za EST (“expressed sequence tags”) sekvence. U interpretaciji podataka, koja je često također dostupna na Internetu, kriju se koji put još ozbiljnije greške. Na kraju je izvorna literatura ipak najbolji pokazatelj, koliko su podaci koje želimo upotrebiti solidni, pa zato sve banke podataka citiraju i izvornu literaturu. Kako je danas i potpuni tekst znanstvenih članaka publiciranih u zadnjih par godina dostupan svakome preko Interneta, nema opravdanja da se to ne pročita prije nego što se “uleti” u neko istraživanje iz bioinformatike.

Raspon istraživanja u bioinformatici zaista je ogroman. Moja radna grupa već više godina radi na razvoju alata za analizu membranskih proteina, a tom problemu, koliko god da je važan, udžbenici iz bioinformatike posvećuju samo jednu do dvije strane (1-4). Temeljna pitanja koja financiraju istraživanja u bioinformatici postavljaju se već dugo vremena kao pitanja iz medicine, farmakologije, i primjenjene biologije: Kako liječiti bolesti? Kako spriječiti infekciju? Kako proizvesti dovoljno hrane za čitavo čovječanstvo. Ukratko, cilj moderne biologije je da uz pomoć bioinformatike identificira sve gene u svim živim bićima, da odredi strukturu i funkciju pripadnih proteina, da shvati kako regulacija i interakcija proteina dovodi do naizgled razumnog ponašanja svake, pa i najmanje stanice i do vrlo kompleksnog ponašanja živih bića kao jedinki i ekoloških sustava. Zavodi, instituti i odjeli za bioinformatiku pri sveučilištima i pod okriljem farmaceutskih kompanija, niču u zadnjih par godina kao gljive poslije kiše. Ti trendovi još nisu sasvim stigli do Hrvatske.

Za korisnike bioinformatičkih usluga važno je razumijevanje da se veliki dio biologije i biokemije eliminira kada se prave bioinformatički modeli. U stvarnosti su DNK i proteini komplicirane trodimenzionalne makromolekule koje međudjeluju i sa drugim takvim molekulama i sa malim molekulama okoline u kojoj se nalaze. Sve se to zanemaruje kada se kao polazište za bioinformatičke analize uzme linearni niz (sekvencija) nukleotida u DNK, odnosno aminokiselina u polipeptidnom lancu proteina. Ta radikalna redukcija na jednodimenzionalne (1D) sekvencije jedna je od najplodnijih ideja modeliranja u bioinformatici. To je omogućilo da se računalske metode iz informatike, neke razvijene baš za proučavanje jezika, primijene i u bioinformatici. Jedno vrlo jednostavno i plodno pitanje, koje se može postaviti o dva niza slova, je da li se ti nizovi podudaraju, a ako se ne podudaraju koliko su slični. Usporedba para DNK, ili para polipeptidnih sekvencija danas je temeljna tehnika u biologiji koju svaki istraživač mora znati primijeniti bilo da je bioinformatičar ili ne. Rezultati takvih usporedba omogućuju da se genima pridruži biološka funkcija i da se informacija pridružena jednom genomu prenosi na druge genome. Rezultati velikog broja takvih usporedba omogućuju i da se dođe do dalekosežnih zaključaka o evoluciji gena i genoma. Naime, funkcionalno važne sekvencije imaju tendenciju da su puno bolje sačuvane u tijeku

evolucije. Kvantitativne razlike ili sličnosti u genima danas su za istraživače čak i važnije od kvalitativnih anatomskih opisa.

Bioinformatičari su tako otkrili da su homologni geni najčešće međusobno povezani baš preko evolucijskog procesa. Ortologni geni (“orthologs”) su homologni geni koji dijele istu funkciju, makar se danas mogu naći u vrlo različitim organizmima kao što su čovjek, miš, vinska mušica i gljivica kvasca. Paralogni geni (“paralogs”) su homologni geni koji su nastali duplikacijom gena u istoj vrsti živih bića, no danas više nemaju istu funkcionalnu ulogu. Priroda je vrlo konzervativna. Kada napravi novu vrstu živih bića, ne pravi pri tome i novu biokemiju, nego uglavnom koristi male promjene u već postojećim genima. Kada istoj vrsti daje sasvim nove sposobnosti (kao što je recimo razvoj mozga kod ljudi ili centra za govor u mozgu) vjerujemo da je to također u stanju postići pomoću malih konzervativnih promjena u genima i u regulaciji aktivnosti gena koje majmuni dijele s nama. Dakle, nemamo razloga očekivati da kao inteligentnija vrsta imamo više gena od recimo čimpanza.

No kako naći gene u DNK sekvenciji? Bilo koja DNK sekvencija može se prevesti na čak šest raznih načina u polipeptidnu sekvenciju. Sekvencija se može prevoditi s lijeva na desno ili obrnuto. Nadalje, pošto svaku aminokiselinu u proteinu specificiraju tri nukleotida, postoje tri moguća prevoda za DNK sekvenciju u izabranom smjeru. Prvi prevod počinje s prvim slovom (nukleotidom) u sekvenciji, drugi s drugim, a treći s trećim slovom. U žargonu bioinformatičara ORF (“open reading frame”) bez signala za zaustavljanje sinteze polipeptidnog lanca (stop kodon) dobar je kandidat za gen ili dio gena. ORF regije u DNK sekvenciji imaju statistički drugačiju distribuciju nukleotida, od ostalih DNK regija, pa to koriste svi bioinformatički alati za pronalaženje gena. Neki od tih alata su recimo GenMark (Borodovsky i McIninch, 1993), GAIL (Xu, 1994), PROCRUSTES (Gelfand i Pevzner, 1996), GENSCAN (Burge i Karlin, 1997). Pogreške u nalaženju gena još su vrlo česte, pa ih neki algoritmi, kao na primjer PROCRUSTES smanjuju tako da prevode potencijalne ORF u polipeptidnu sekvenciju i onda traže da li već postoje takve homologne sekvencije. Ako je homologija dovoljno velika, vjerojatnost da se to desi slučajno vrlo je mala, pa to znači da potencijalni ORF vjerojatno kodira za pravi gen.

Prvi popularni program za određivanje homologije nazvan je FASTA (Lipman i Pearson, 1985). Sa ALIGN programom mogu se sve sekvencije u FASTA formatu automatski staviti jedna ispod druge tako da što veći broj identičnih aminokiselina sravnjeno tako da dođu točno jedna ispod druge. Tako se naravno lakše vidi koje su aminokiseline sačuvane (pretpostavljamo da su te aminokiseline funkcionalno važnije) u tijeku evolucije. Danas je popularniji program BLAST (“Basic Local Alignment Tool”), koji traži područja lokalne sličnosti između dvije sekvencije. BLAST je tako brz da ga istraživači rutinski koriste za pretragu svih DNK sekvenci u GenBank ili svih proteinskih sekvenci u Swiss-Prot bazi podataka, da bi otkrili da li njihova sekvencija ima

negdje svog homolognog para. Više informacija o BLAST-u i o drugim bioinformatičkim alatima može se naći na adresi: <http://ncbi.nlm.nih.gov>. Jedna prednost BLAST-a pred ostalim sličnim programima je da BLAST daje i informaciju koliko je nađena sličnost između sekvencija značajna. Velika E-vrijednost, veća od 5 ili 10 znači da nađena sličnost među sekvencijama može biti i slučajna. Mala E-vrijednost, manja od 0.1 ili 0.05 znači da je nađena sličnost značajna. Grubo empirijsko pravilo kaže da sličnost manja od 25% između dvije sekvencije znači da vjerojatno nisu evolucijski povezane. Sličnost do 15% nalazi se i između dvije potpuno slučajne sekvencije. Sličnost znači da postoje identična i slična mjesta u dvije sekvencije. Recimo leucin i izoleucin su slične hidrofobne aminokiseline koje obično mogu zamijeniti jedna drugu bez da se funkcija ili struktura proteina zbog toga bitnije promijeni.

Istovremena usporedba većeg broja sekvenci najčešće se radi na proteinskim sekvencijama. Vrlo je popularan algoritam ClustalW (Higgins, 1996), koji sravnjuje sekvencije jednu ispod druge tako da se identične aminokiseline na nekom rednom mjestu nalaze točno jedna ispod druge. ClustalW radi i višestruka sravnjivanja (za veći broj sekvencija) i to ne samo prema identitetu nego i prema sličnosti aminokiselina. Na osnovi rezultata koji se dobiju sa ClustalW ili ClustalX programima moguće je konstruirati filogenetsko stablo koje daje dodatne informacije o evoluciji određenog gena. Najpopularniji specijalizirani program koji to radi je PHYLIP (Felsenstein, 1996). PHYLIP je zapravo paket od 30 programa za filogenetsku analizu. Može se instalirati na velikom broju različitih operativnih sustava od DOS-a do UNIX-a. Rezultati najviše ovise o tome koliko je ulazna datoteka dobro pripremljena (koliko je dobro sravnjavanje). Takva su istraživanja pokazala da neki proteini evoluiraju brže od drugih u istom genomu. S druge strane isti ili vrlo slični protein brže evoluiraju u jednim nego u drugim organizmima. Zbog toga nije lako povezati rezultate filogenetske analize s paleontološkim podacima o vjerojatnom povijesnom periodu kada je došlo da odvajanja vrsta.

Kod usporedbe sekvencija od vrlo su velike koristi motivi za koje se zna da su karakteristični za određene obitelji proteina. Tako Swiss Institute of Bioinformatics održava ne samo Swiss-Prot datoteku dobro opisanih proteinskih sekvencija, nego i PROSITE datoteku sa preko 1200 karakterističnih motiva. Motivi koje PROSITE sadrži skupljeni su iz publiciranih članaka.

## **Proteomika i SPLIT**

Iz fokusa na genomiku bioinformatika sve više prelazi na proteomiku. Proteomika želi doznati 3D strukturu za sve proteinske obitelji koje postoje na Zemlji. Za sada je u Protein Data Bank (PDB) skupljeno više od 20000 3D struktura, no samo je nešto više od 1000 potpuno različitih

nehomolognih struktura. Skupiti što više novih različitih struktura najvažniji je cilj današnje bioinformatike, jer će na osnovu poznavanja svih glavnih motiva savijanja proteina u 3D strukturu biti moguće predvidjeti 3D strukturu za većinu novih dekodiranih sekvencija.

U mojoj radnoj grupi napravili smo alate koje smo nazvali SPLIT i GARLIC, od kojih prvi služi za analizu primarne strukture membranskih proteina, a drugi za vizualizaciju makromolekula poznate strukture. Oba su alata slobodno dostupna svima na adresi: <http://garlic.mefos.hr>. Naš poslužitelj SPLIT, koji je fizički lociran u Osijeku od 1998-me godine, prvi je znanstveni poslužitelj uopće u Hrvatskoj i naravno prvi poslužitelj kod nas za proračune iz bioinformatike. Kao i mnogi drugi alati za bioinformatičke analize GARLIC zahtijeva UNIX ili LINUX operativni sustav. Ozbiljna istraživanja u bioinformatici ne mogu se ograničiti samo na upotrebu Windowsa kao operativnog sustava. Međutim većina krajnjih korisnika bioinformatičkih alata, još uvijek koristi Windowse, tako da je i velik broj tih alata prilagođen za rad u Windows okolini. To je slučaj i sa SPLIT-om, koji radi pod tri operativna sustava: DOS, Windows i LINUX. SPLIT je do sada koristilo tisuće laboratorija iz svijeta, najviše iz zemalja kao što su SAD, Engleska i Njemačka koje imaju vrlo razvijene biološke znanosti. Za “on-line” web poslužitelje, poput SPLIT-a, bitno je da su prijateljski raspoređeni prema korisniku, koji ne treba učiniti ništa drugo nego prilijepiti svoju sekvenciju u zadani okvir i nakon što pritisne dugme “submit” čekati minutu ili manje da se pojavi grafički rezultat analize. Izvorni kod za SPLIT u verziji koja radi pod LINUX-om ili pod Windowsom može slobodno svatko preuzeti. Oni koji to učine većinom su najozbiljniji istraživači, koji nisu zadovoljni da mogu analizirati samo jednu proteinsku sekvenciju odjednom, niti mogu prihvatiti da je taj program za njih crna kutija koja samo ne-zna-se-kako izbacuje krajnje rezultate.

Naš program GARLIC (Zucić, 2001) za vizualizaciju 3D struktura jedan je od najboljih programa te vrste. Radi pod LINUX operativnim sustavom i već je uvršten u nekoliko standardnih kolekcija programa. Vrlo popularan je i RasMol (Bernstein), Cn3D (NCBI), PyMol i SWISS-PDB Viewer. MolScript se često koristi za prikazivanje 3D strukture za publikaciju. Posebni programi koriste se za pridruživanje sekundarne strukture polipeptidnom lancu na osnovu poznate 3D strukture. To su na primjer DSSP i STRIDE. Baza podataka SCOP grupira sve PDB strukture prema njihovoj sekundarnoj strukturi kao alfa, beta i razne miješane alfa-beta strukture. Sekundarna i terciarna struktura proteina često su čak bolje sačuvane od njihove primarne strukture. To je i razumljivo, jer tek 3D struktura postaje enzimatski aktivna.

Predviđanje strukture proteina na osnovu sekvencije najvažniji je i najteži cilj u bioinformatici. Za 99% sekvencija ne postoji 3D struktura, pa teorijske metode predviđanja 2D i 3D strukture ostaju kao jedina mogućnost da se bar donekle sazna struktura koja je temelj za proučavanje djelovanja proteina. Osim toga, određivanje 3D strukture pomoću difrakcije s X zrakama ili pomoću NMR tehnike, posao je za veliki tim istraživača koji imaju dugogodišnje takvo

iskustvo, pa je zbog toga to skupo (u prosjeku 100000\$ po proteinu), dugotrajno, a ponekad i nemoguće. Kristalizacija proteina u praksi je preduvjet za upotrebu X-zraka radi određivanja 3D strukture, međutim dobiti pravilne kristale dovoljne veličine (oko 1 mm) još uvijek je više umjetnost nego znanost. To je posebno slučaj s membranskim proteinima kojih je od 20 do 30% od svih proteina, a do danas su istraživači uspjeli dobiti 3D strukturu samo za 20 do 30 takvih proteina. Za takve proteine, teorijsko predviđanje strukture na osnovu jeftino dobivene sekvencije iz projekta genoma, nije samo jeftinija nego je i jedina alternativa.

Postoje dva prilaza predviđanju 3D strukture. Češći prilaz je da se iz baze podataka postojećih struktura izdvoje optimalne strukture pomoću kojih se modelira zadana sekvencija, ili pak da se iz PDB struktura izluče parametri, poput sklonosti za određenu sekundarnu strukturu, koji poboljšavaju predviđanje strukture za zadanu sekvenciju. Drugi prilaz osniva se na simulaciji fizikalnih sila i molekularne dinamike. Najbolji algoritmi za predviđanje 3D strukture, poput Modeller (<http://guitar.rockefeller.edu/modeller/modeller.html>), koriste i jedan i drugi prilaz. Modeller je najvećim dijelom konstruirao mladi slovenski znanstvenik Andrej Šali, koji sada vrlo uspješno radi u SAD.

Predviđanje 3D strukture izravno iz primarne strukture proteina toliko je težak problem, da ga je dobro pokušati razbiti na manje ambiciozne dijelove. Predviđanje sekundarne strukture često se smatra prvim (lakšim) korakom u predviđanju tercijarne strukture. Iako je moguće definirati veći broj sekundarnih struktura pomoću prije spomenutih algoritama DSSP ili STRIDE, programi predviđanja sekundarne strukture većinom koriste model u kojem postoje samo tri ili samo četiri strukture. To su alfa-uzvojnica, beta nabrana ploha ili beta nit, zavoj i sve drugo što se svrstava pod nepoznatu strukturu. Većina modernih alata za predviđanje sekundarne strukture koriste kao ulazne podatke ne samo zadanu sekvenciju nego i njoj homologne sekvencije. To rade PHD i PredictProtein (Rost, 1994), koji postižu točnost predviđanja od 70 do 74%, a također i PSIPRED i Jpred programi koji postižu sličnu točnost predviđanja. Svi najuspješniji programi koriste moderne metode neuralnih mreža ili skrivenih Markovljevih modela. Slaba strana tih metoda je da im se točnost predviđanja smanjuje kada nisu u stanju naći homologne sekvencije zadanoj. Starije metode, koje ne koriste homologiju, imaju točnost predviđanja manju od 70%. No i najveća točnost predviđanja od 75% nije dovoljna da zagantira dobro predviđenu 3D strukturu.

Predviđanje sekundarne strukture membranskih proteina zahtijeva upotrebu za to specijaliziranih algoritama. Obično se prvo predvidi sekventni položaj transmembranskih uzvojnica, jer su proteini s transmembranskim uzvojnica daleko najčešći integralni membranski proteini. Zatim se predvidi topologija membranskog proteina, odnosno usmjerenost transmembranskih uzvojnica i položaj petlji koje ih spajaju na unutrašnjoj ili vanjskoj strani membrane. Algoritam SPLIT radi oba ova zadatka u najnovijoj verziji SPLIT 4.0.



Kao i kod svih drugih poslova u bioinformatici, ako koristite SPLIT pametno je upotrebiti i druge alate za predviđanje strukture membranskih proteina, znati njihove prednosti i slabosti na osnovu podataka iz izvorne literature i naravno potruditi se da doznate sve što se može znati o sekvenciji koja vas zanima. SPLIT daje i adrese/linkove za konkurentske programe i za Swiss-Prot bazu podataka. SPLIT ne koristi homologiju, pa mu se točnost neće smanjiti ako vaša sekvencija kodira za membranski protein i nema mnogo “rođaka”.

Daljnji razvoj SPLIT-a moguć je i poželjan ako se neki mladi istraživač ili znanstveni novak zainteresira za taj problem. U tom algoritmu još nije dobro riješeno pitanje predviđanja signalnih sekvencija na N-kraju proteina, upotreba homologije, treniranje na većoj bazi podataka poznatih struktura i predviđanje sekundarne strukture u petljama koje strše izvan membrane. Korisnik programa SPLIT može izabrati bilo koju od 88 ljestvica aminokiselina, koje već postoje u tom programu, da s njima računa sklonosti nastanka pojedinih sekundarnih struktura i hidrofobne momente. U budućem radu trebalo bi dodati novije ljestvice fizikalno-kemijskih svojstava aminokiselina za koje njihovi autori misle da su bolje. Također nismo zadovoljni ni sa točnošću predviđanja transmembranskih beta niti, koje postoje u nekim membranskim proteinima poput porina. Izvorni kod SPLIT-a napisan je u FORTRAN-u, koji je programski jezik danas u odumiranju, pa bi bilo bolje sve iznova napisati u C ili C++ jeziku. Automatski prevod u C jezik već koristimo prilikom instalacije poslužitelja, koji je fizički lociran u Osijeku. Instalacija zrcalnog ili sličnog poslužitelja u Splitu, još je jedan cilj rada moje grupe, za koju nam treba ekspert u Linux-u ili Unix-u.

Kada i saznamo 3D strukturu svih važnijih proteina u stanici (možda u narednih 20 godina, jer to je mnogo kompleksniji projekt od projekta genoma), još uvijek će ostati izazov da otkrijemo kako je sve to skupa povezano u mrežama biokemijskih reakcija. U zadnje doba moja grupa radi i na tom problemu. Koncentrirali smo se na reakcije bitne u pretvorbi slobodne energije, kao što su reakcije fotosinteze. Dakle područje tog istraživanja spada istodobno i u bioenergetiku (5) i u bioinformatiku. Moguće je modelirati i kompleksne biokemijske reakcije ili pomoću diferencijalnih jednačbi, ili u stacionarnom stanju pomoću algebarskih jednačbi. U stacionarnom stanju koriste se analogije Kirchhoffovih zakona za petlje i čvorišta električne mreže. Istraživači pokušavaju primijeniti i razne principe optimizacije, koji se mogu podijeliti u dvije klase: fizikalni i biološki. Biološki principi su recimo maksimalna enzimatska aktivnost ili maksimalna efikasnost rada, dok su fizikalni principi povezani s pojmom entropije kao mjere za nered u sustavu. Neki istraživači misle da žive stanice teže biti što bliže stanju minimalne produkcije entropije. No život održava svoju kompleksnu strukturu upravo proizvodnjom i eksportom entropije u svoju okolinu. U svakom slučaju, neki općeniti oprincip optimizacije biokemijskih i enzimatskih ciklusa u stanici pomogao bi

da se uvede reda u gustu šumu biokemijskih reakcija, koje određuju ne samo ponašanje stanice, nego i rad organa i čitavog organizma.

## Neke primjene bioinformatike u medicini

Dosadašnji rad sa algoritmom SPLIT pokazao je da se upotrebom SPLIT-a mogu ispraviti mnoge greške u pridruživanju transmembranskih uzvojnica sekventnim lokacijama membranskih proteina iz Swiss-Prot datoteke (“annotation errors”). To je važno i za medicinu, jer mnogi membranski receptori i voltažni kanali igraju važnu ulogu u radu mozga, srca i drugih organa, pa je potrebno znati gdje se nalaze funkcionalno važna mjesta, da li u citoplazmi ili izvan stanice. Naravno, stanična lokacija takvih mjesta ovisi o broju i usmjerenju transmembranskih uzvojnica.

Na žalost, pomaci u otkrivanju uzroka bolesti daleko su brži nego pomaci u otkrivanju najbolje metode liječenja. No jedno bez drugoga ne ide. Moramo što bolje znati molekularne uzroke bolesnih stanja, da bi primijenili racionalne metode liječenja. Za sada, podaci o genima u kojima defekti izazivaju pojavu genetičkih bolesti, više koriste molekularnim biolozima da upoznaju mehanizam djelovanja pripadnih proteina, nego samim bolesnicima. Nadamo se da će se ta situacija u budućnosti promijeniti.

U prilogu je spisak gena i pripadnih membranskih proteina u kojima defekti izazivaju pojavu poznatijih nasljednih bolesti, koju smo pripremili u našem laboratoriju (Tablica 1). Odmah se vidi da često nema obostrano jednoznačne korespodencije između pojedinih gena i pojedinih bolesti (samo važnije bolesti pridružene istoj sekvenciji spominju se u Tablici 1). Ista bolest može nastati zbog mutacija u više raznih gena, a mutacije u istom genu mogu uzrokovati pojavu različitih bolesti. Važna uloga membranskih proteina u transportu iona kroz membranu, prenosu signala, i izgradnji višestaničnog organizma, također je očigledna iz Tablice 1. Poznavanje lokacije funkcionalno važnih dijelova membranskog proteina – u membrani, u citoplazmi ili izvan citoplazme i te kako pomaže u medicini i farmakologiji, a to je upravo ono što nam daje ekspertni sustav poput SPLIT-a. Da bi predvidio lokaciju segmenata, odnosno transmembransku topologiju, SPLIT prvo daje vrlo točno predviđanje broja potencijalnih transmembranskih uzvojnica (TMU). Iz Tablice 1 vidi se da je moguće napraviti grubu podjelu na klase membranskih proteina slične strukture i funkcije, već na osnovu očekivanog broja TMU, a naravno još bolje kada se doda i podatak o usmjerenosti svake TMU. Za dosta sekvenciji nismo još sigurni koliko im treba pridružiti TMU i pogotovo kakva im je točno topologija u membrani. Dodali smo upitnik u trećem stupcu Tablice 1, gdje po našem mišljenju postoji takva nesigurnost, ili je nedavno došlo do promjene u broju očekivanih TMU. Oznaka  $\frac{1}{2}$  u zadnjem retku tog stupca odnosi se na specijalan slučaj TMU koja se vjerojatno savija poput ukosnice, tako da izlazi na istu stranu membrane gdje je i ušla i

možda ne prodire više nego u jedan sloj fosfolipidnog dvosloja membrane. Dodatni podaci o molekularnim uzrocima nasljednih bolesti mogu se naći u bazi podataka OMIM (Online Mendelian Inheritance in Man) na adresi: <http://www.ncbi.nlm.nih.gov/Omim>. OMIM (McKusick) pruža obilje podataka o svim patološkim stanjima koja imaju izvor u genetici.

I na kraju dužni smo odgovoriti na pitanje koji glicin izaziva pojavu nasljedne gluhoće kada se pretvori u serin. Iskoristimo FASTA ili BLAST iz citirane kolekcije bioinformatičkih alata da izvršimo pretragu proteinske baze podataka na osnovu sličnosti zadanog segmenta. Prilijepimo segment DSLWWGTITLTTIGYGDKTP u za to predviđeni prozorčić i obavimo tu pretragu bilo sa FASTA ili sa BLAST alatom. Oba alata brzo nalaze da protein CIQ4\_HUMAN iz Swiss-Prot baze podataka ima taj segment u svojoj primarnoj strukturi. Kada otvorimo datoteku CIQ4\_HUMAN u Swiss-Prot bazi (<http://us.expasy.org/cgi-bin/niceprot.pl?P56696>) vidimo da je taj protein voltažni kanal za kalij posebno važan za ispravni rad slušnog organa. Zadani segment upravo je segment pore koja selektivno propušta samo kalij, a mutacije G(285)S ili G(285)C iz tog segmenta izazivaju pojavu nasljedne gluhoće. Glicin 285 prvi je glicin iz tripleta GYG. Iz literature, koristeći recimo PubMed bazu podataka (i naziv pripadnog gena, koji možemo naći i u Tablici 1.), možemo saznati da je GYG zajednički motiv unutar segmenta pore i za mnoge druge voltažne kanale za kalij, pa se smatra dijelom selekcijskog filtera specifičnog za propuštanje iona kalija. Algoritam SPLIT daje nam naslutiti i sekundarnu strukturu zadanog segmenta. Njegov N-kraj mogla bi biti kratka u membranu uronjena uzvojnica, koja prema C-kraju pravi zavoje i pretvara se u izduženu strukturu beta niti. Beta nit izbija ponovo na istu stranu membrane tako da čitav segment pore mora imati oblik petlje ili ukosnice. Detaljniji modeli pokazuju da selekcijski filter GYG pravi najuži dio pore u strukturi beta niti, a strukturu beta-niti za GYG predviđa i SPLIT. Zašto se pak defekt u propuštanje kalijevih iona kroz membranu pojavljuje na razini slušnog organa kao gluhoća, posebna je fascinantna tema koja izlazi izvan okvira ovog članka.

## Literaturna vrela:

- 1) Cynthia Gibas i Per Jambeck: "Developing Bioinformatics Computer Skills". O'Reilly&Associates, Cambridge 2001.
- 2) Andreas Baxevanis i Francis Ouellette: "Bioinformatics. A Practical Guide to the Analysis of Genes and Proteins". Wiley, New York, 1998.
- 3) Des Higgins i Willie Taylor: "Bioinformatics: Sequence, Structure and Databanks: A Practical Approach", Oxford Univ. Press, Oxford, UK, 2000
- 4) Arthur M. Lesk: "Introduction to Bioinformatics", Oxford Univ. Press, Oxford, UK, 2002
- 5) Davor Juretić: "Bioenergetika, rad membranskih proteina". Informator, Zagreb, 1997.

## Baze podataka

SWISS PROT <http://us.expasy.org/sprot/>

EMBL <http://www.ebi.ac.uk/embl/>

PROSITE <http://us.expasy.org/prosite/>

PDB- Protein Data Bank <http://www.rcsb.org/pdb/>

OMIM <http://www3.ncbi.nlm.nih.gov/omim/>

## Alati

Pretraživanje proteinskih baza podataka na osnovu sličnosti primarnog slijeda

Blast <http://us.expasy.org/tools/blast/>

Fasta <http://www.ebi.ac.uk/fasta33/>

Sravnjavanje sekvenca

CLUSTALW <http://www.ch.embnet.org/software/ClustalW.html>

Predviđanje sekundarne strukture

[PredictProtein](#) - PHDsec, PHDacc, PHDhtm, PHDtopology, PHDthreader, MaxHom, EvalSec --  
Columbia University <http://cubic.bioc.columbia.edu/predictprotein/>

[Jpred](#) – Kombinacija nekoliko metoda za predviđanje sekundarne strukture proteina - EBI  
<http://www.compbio.dundee.ac.uk/~www-jpred/>

[nnPredict](#) - University of California at San Francisco (UCSF)  
<http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html>

[PSIpred](#) – Razne metode za predviđanje strukture proteina - Brunel University  
<http://bioinf.cs.ucl.ac.uk/psipred/>

Predviđanje topologije membranskih proteina i vizualizacija 3D structure makromolekula

Davor Juretić radna grupa –SPLIT, GARLIC – <http://garlic.mefos.hr>

Pretraživanje literature: PubMed – <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>

Tablica 1. Membranski proteini i nasljedne bolesti (za lakše snalaženje po Internetu korišteni su engleski termini za genetičke bolesti). Bolesti su sređene prema broju transmembranskih uzvojnica (#TMU) pripadnog polipeptida i broju poznatih mutacija (# MUT )(početkom 2003. godine) u aminokiselinama (AA) koje uzrokuju tu bolest.

| Protein (gen)<br>Swiss-Prot kod | Genetička bolest                                  | #<br>TMU | #<br>MUT | #<br>AA | Djelovanje/<br>funkcija                |
|---------------------------------|---------------------------------------------------|----------|----------|---------|----------------------------------------|
| ccaa(cacn1a4)                   | Episodic ataxia; Familial hemiplegic migraine     | 24       | 4        | 2505    | Ca <sup>++</sup> kanal                 |
| ccaf(cacna1f)                   | Night blindness                                   | 24       | 4        | 1966    | Ca <sup>++</sup> kanal                 |
| ccas(cacna1s)                   | Hypokalemic periodic paralysis                    | 24       | 4        | 1873    | Ca <sup>++</sup> kanal                 |
| cin1(scn1a)                     | Generalized epilepsy; Myoclonic epilepsy          | 24       | 8        | 2009    | Na <sup>+</sup> kanal                  |
| cin4(scn4a)                     | Hyperkalemic periodic paralysis                   | 24       | 13       | 1836    | Na <sup>+</sup> kanal                  |
| cin5(scn5a)                     | Long QT syndrome 3, Cardiac arrhythmia            | 24       | 13       | 2016    | Na <sup>+</sup> kanal                  |
| acc8(kcnj11)                    | Familial persistent hyperinsulinemic hypoglycemia | 17 ?     | 22       | 1580    | K <sup>+</sup> kanal                   |
| mrp6(abcc6)                     | Pseudoxanthoma elasticum                          | 17 ?     | 25       | 1503    | Anioni transp.                         |
| abcr(abcr)                      | Stargardt disease                                 | 14 ?     | 198      | 2273    | Transporter                            |
| abc1(abca1)                     | Tangier disease                                   | 13 ?     | 20       | 2261    | Anioni transp.                         |
| cicl(clcnkb)                    | Bartter syndrome 3                                | 13       | 5        | 687     | Cl <sup>-</sup> kanal                  |
| clc1(clcn1)                     | Becker's dis.; Thomsen's dis.; Myotonia levior    | 13       | 30       | 988     | Cl <sup>-</sup> kanal                  |
| sl55(slca5a5)                   | Iodide transport defect                           | 13 ?     | 5        | 643     | Na-I transp.                           |
| b3at(slca4a1)                   | Hereditary spherocytosis                          | 12       | 34       | 911     | Anioni kanal                           |
| cftr(cftr)                      | Cystic fibrosis; Male sterility                   | 12       | 146      | 1480    | Cl <sup>-</sup> transport              |
| dtd(dtd)                        | Diastrophic dysplasia                             | 12       | 6        | 739     | SO <sub>4</sub> <sup>-</sup> transport |
| p(p)                            | Albinism                                          | 12       | 26       | 838     | Tyr transport                          |
| rhag(rhag)                      | Chronic hemolytic anemia                          | 12       | 1        | 409     | Transport ?                            |
| ryr1(ryr1)                      | Malignant hyperthermia; Central core disease      | 12       | 26       | 5038    | Ca <sup>++</sup> kanal                 |
| ryr2(ryr2)                      | Arrhythmogenic cardiomyopathy; Tachycardia        | 12       | 9        | 4967    | Ca <sup>++</sup> kanal                 |
| s123(slca12a3)                  | Gitelman syndrome                                 | 12       | 14       | 1021    | Na-Cl transp.                          |
| s6a2(slca6a2)                   | Orthostatic intolerance                           | 12       | 1        | 617     | Amine transp.                          |
| tht1(slca19a2)                  | Thiamine-responsive megaloblastic anemia          | 12       | 3        | 497     | Tiamin transp                          |
| yla1(slca7a7)                   | Lysinuric protein intolerance                     | 12 ?     | 4        | 511     | AA transport                           |
| npc1(npc1)                      | Niemann-Pick disease type 2                       | 11 ?     | 22       | 1278    | Kolesterol tr.                         |
| ocn2(slca22a5)                  | Carnitine deficiency                              | 11 ?     | 7        | 557     | Karnitin trasp.                        |
| pend(pds)                       | Pendred syndrome                                  | 11 ?     | 12       | 780     | Cl, I transport                        |

|               |                                                                   |      |    |      |                             |
|---------------|-------------------------------------------------------------------|------|----|------|-----------------------------|
| pkd1(pkdl)    | Polycystic kidney disease type 1                                  | 11 ? | 62 | 4303 | Anioni kanal                |
| sl51(slc5a1)  | Glucose-galactose malabsorption                                   | 11   | 2  | 664  | Na <sup>+</sup> -Glukoza t. |
| wfs1(wfs1)    | Wolfram syndrome                                                  | 11 ? | 25 | 890  | ?                           |
| a8b1(atp8b1)  | Intrahepatic cholestasis                                          | 10   | 6  | 1251 | PL transport                |
| ata2(tp2a2)   | Darier's disease                                                  | 10   | 33 | 1042 | Ca <sup>++</sup> transport  |
| atc1(atp2c1)  | Hailey-Hailey disease                                             | 10   | 9  | 919  | Ca <sup>++</sup> transport  |
| clc5(clcn5)   | Dent's disease                                                    | 10 ? | 9  | 746  | Cl <sup>-</sup> kanal       |
| xk(xk)        | McLeod syndrome                                                   | 10 ? | 2  | 444  | Transport                   |
| at7a(atp7a)   | Menkes syndrome                                                   | 8    | 14 | 1500 | Cu <sup>++</sup> transport  |
| at7b(atp7b)   | Wilson disease                                                    | 8 ?  | 98 | 1465 | Cu <sup>++</sup> transport  |
| cln3(cln3)    | Batten disease                                                    | 8 ?  | 6  | 438  | ?                           |
| ntci(slc10a2) | Bile acid malabsorption, Crohn's disease                          | 8?   | 3  | 348  | Žučne kis. tr.              |
| psn1(psen1)   | Alzheimer disease 3                                               | 8 ?  | 62 | 467  | ?                           |
| actr(acthr)   | Glucocorticoid deficiency                                         | 7    | 8  | 297  | Receptor                    |
| casr(casr)    | Hypocalciuric hypercalcemia                                       | 7    | 16 | 1078 | Ca <sup>++</sup> senzor     |
| etbr(ednrb)   | Hirschsprung disease; Shah-Waardenburg synd.                      | 7    | 7  | 442  | Receptor                    |
| hh3r(hrh3)    | Shy-Drager syndrome                                               | 7    | 1  | 445  | Receptor                    |
| lshr(lhcgr)   | Male pseudohermaphroditism; Leydig cell hypoplasia                | 7    | 24 | 699  | Receptor                    |
| mc4r(mc4r)    | Obesity                                                           | 7    | 6  | 332  | Receptor                    |
| oa1(oa1)      | Ocular albinism type 1                                            | 7 ?  | 13 | 404  | ?                           |
| opsb(opn1sw)  | Tritan color blindness                                            | 7    | 3  | 348  | Photoreceptor               |
| opsd(rho)     | Retinitis pigmentosa; Stationary night blindness                  | 7    | 69 | 348  | Photoreceptor               |
| opsg(opn1mw)  | Deutan color blindness                                            | 7    | 1  | 364  | Photoreceptor               |
| opsr(opn1lw)  | Protan color blindness                                            | 7    | 2  | 364  | Photoreceptor               |
| tshr(tshr)    | Graves' disease (hyperthyroidism); Thyroid adenomas and neoplasms | 7    | 48 | 764  | Receptor                    |
| v2r(avppr2)   | Diabetes insipidus, nephrogenic                                   | 7    | 30 | 371  | Receptor                    |
| aqp2(aqp2)    | Diabetes insipidus, nephrogenic                                   | 6    | 4  | 271  | H <sub>2</sub> O kanal      |
| atp6(atp6)    | Leigh syndrome; Leber optic neuropathy                            | 6 ?  | 4  | 226  | H <sup>+</sup> kanal        |
| c24b(czbb)    | Chronic granulomatous disease                                     | 6 ?  | 37 | 569  | e <sup>-</sup> transport    |
| cik1(kcna1)   | Episodic ataxia with myokymia                                     | 6    | 7  | 495  | K <sup>+</sup> kanal        |
| ciq1(kcnq1)   | Long QT synd. 1; Jervell&Lange-Nielsen disease                    | 6    | 72 | 676  | K <sup>+</sup> kanal        |
| ciq2(kcnq2)   | Epilepsy benign neonatal type 1                                   | 6    | 4  | 872  | K <sup>+</sup> kanal        |

|                |                                                  |     |     |      |                        |
|----------------|--------------------------------------------------|-----|-----|------|------------------------|
| ciq3(kcnq3)    | Epilepsy benign neonatal type 2                  | 6   | 2   | 872  | K <sup>+</sup> kanal   |
| ciq4(kcnq4)    | Deafness, autosomal dominant                     | 6   | 6   | 695  | K <sup>+</sup> kanal   |
| g6pt(g6pt)     | Glycogen storage (von Gierke)                    | 6   | 24  | 357  | ?                      |
| kch2(kcnh2)    | Long QT syndrome type 2                          | 6   | 65  | 1169 | K <sup>+</sup> kanal   |
| mip(mip)       | Cataract                                         | 6   | 2   | 263  | H <sub>2</sub> O kanal |
| ort1(slc25a15) | Hyperammonemia                                   | 6 ? | 7   | 301  | Ornitin transp.        |
| pkd2(pkcd2)    | Polycystic kidney disease type 2                 | 6 ? | 5   | 968  | Ca <sup>++</sup> kanal |
| tmc1(tmc1)     | Deafness 36; Deafness 7                          | 6 ? | 2   | 760  | ?                      |
| ucp3(ucp3)     | Obesity severe                                   | 6 ? | 2   | 312  | H <sup>+</sup> kanal   |
| ach4(chrna4)   | Epilepsy nocturnal frontal lobe 1                | 4   | 1   | 627  | Recep.& kanal          |
| achb(chrnb1)   | Slow-channel myastenic syndrome                  | 4   | 2   | 501  | Recep.& kanal          |
| ache(chrne)    | Myastenic syndrome                               | 4   | 6   | 493  | Recep.& kanal          |
| cxbl(gjb1)     | Charcot-Marie-Tooth disease                      | 4   | 137 | 283  | Kanal                  |
| cxbl(gjb2)     | Deafness; Vohwinkel's syndrome                   | 4 ? | 9   | 226  | Kanal                  |
| cxbl(gjb3)     | Deafness; Erythrokeratoderma variabilis          | 4 ? | 8   | 270  | Kanal                  |
| glra1(glra1)   | Startle disease (hyperekplexia)                  | 4   | 9   | 449  | Recep.& kanal          |
| mypr(plp)      | Pelizaeus-Merzbacher disease; Spastic paraplegia | 4   | 59  | 276  | Adhezija               |
| pm22(pmp22)    | Charcot-Marie-Tooth and Dejerine-Sottas syndr.   | 4   | 23  | 160  | ?                      |
| rds(rds)       | Retinitis pigmentosa                             | 4   | 41  | 346  | Adhezija               |
| irk1(kcnj1)    | Bartter syndrome antenatal                       | 2   | 9   | 391  | K <sup>+</sup> kanal   |
| irkb(kcnj11)   | Hyperinsulinemic hypoglycemia of infancy         | 2   | 3   | 390  | K <sup>+</sup> kanal   |
| prf1(prf1)     | Hemophagocytic lymphohistiocytosis type 2        | 2   | 8   | 555  | Kanal                  |
| scab(scnn1b)   | Liddle syndrome (hypertension)                   | 2   | 3   | 640  | Na <sup>+</sup> kanal  |
| ush3(ush3a)    | Usher syndrome type 3                            | 2 ? | 2   | 120  | ?                      |
| a4(app)        | Alzheimer's disease; haemorrhaging + amyloidosis | 1   | 6   | 770  | Receptor               |
| caml(caml1)    | Hydrocephalus; Spastic paraplegia; Masa synd.    | 1   | 29  | 1257 | Adhezija               |
| cygd(gucy2d)   | Leber congenital amaurosis; Cone-rod dystrophy   | 1   | 11  | 1103 | cGMP sinteza           |
| cyrg(il2rg)    | Immunodeficiency severe combined                 | 1   | 35  | 369  | Receptor               |
| dcc(dcc)       | Colorectal tumors                                | 1   | 2   | 1447 | Adhezija?              |
| dha4(alhd3a2)  | Sjogren-Larsson syndrome                         | 1   | 7   | 485  | FA sinteza             |
| egln(eng)      | Rendu-Osler-Weber syndrome                       | 1   | 7   | 658  | Adhezija               |
| fgr1(fgfr1)    | Pfeiffer syndrome                                | 1   | 1   | 822  | Recept. Tyr k.         |
| fgr2(fgfr2)    | Crouzon synd.; Jackson-Weiss synd.; Apert synd.  | 1   | 36  | 821  | Recept. Tyr k.         |
| fgr3(fgfr3)    | Achondroplasia; Thanatophoric dysplasia          | 1   | 14  | 806  | Recept. Tyr k.         |

|              |                                                  |     |     |      |                      |
|--------------|--------------------------------------------------|-----|-----|------|----------------------|
| ghr(ghr)     | Laron-type pituitary dwarfism                    | 1   | 9   | 638  | Receptor             |
| gpba(gp1ba)  | Von Willebrand disease; Bernard-Soulier synd.    | 1   | 10  | 626  | Adhezija             |
| il4l(il4l)   | Asthma atopic                                    | 1   | 2   | 825  | Receptor             |
| insr(insr)   | Diabetes mellitus; Leprechaunism                 | 1   | 22  | 1382 | Receptor             |
| itb2(itgb2)  | Leukocyte adhesion deficiency                    | 1   | 11  | 769  | Adhezija             |
| itb4(itgb4)  | Epidermolysis bullosa                            | 1   | 14  | 1822 | Adhezija             |
| kir3(acvrl1) | Hemorrhagic telangiectasia                       | 1   | 11  | 503  | Receptor             |
| kit(kit)     | Piebaldism; Mast cell leukemia                   | 1   | 7   | 976  | Recept. Tyr k.       |
| ldlr(ldlr)   | Familial hypercholesterolemia                    | 1   | 138 | 860  | Receptor             |
| met(met)     | Renal cell carcinoma                             | 1   | 9   | 1390 | Recept. Tyr k.       |
| mink(kcne1)  | Jervell and Lange-Nielsen; Long QT syndrome 5    | 1   | 11  | 129  | K <sup>+</sup> kanal |
| mir1(kcne2)  | Long QT syndrome type 6                          | 1   | 4   | 123  | K <sup>+</sup> kanal |
| mir2(kcne3)  | Hypokalemic periodic paralysis                   | 1   | 1   | 103  | K <sup>+</sup> kanal |
| myp0(mpz)    | Charccot-Marie-Tooth disease                     | 1   | 45  | 248  | Adhezija             |
| ntc3(notch3) | Arteriopathy with subcortical insults (CADASIL)  | 1 ? | 25  | 2321 | Receptor             |
| pexd(pex13)  | Adrenoleukodystrophy (peroxisome disorder)       | 1   | 1   | 403  | Protein transp.      |
| piga(piga)   | Paroxysmal nocturnal hemoglobinuria              | 1   | 2   | 484  | PI sinteza           |
| ret(ret)     | Multiple endocrine neoplasia; Hirschprung dis.   | 1   | 88  | 1114 | Recept. Tyr k.       |
| trka(ntrk1)  | Cipa disease (insensitivity to pain, anhidrosis) | 1   | 11  | 796  | Recept. Tyr k.       |
| tyro(tyr)    | Albinism                                         | 1   | 59  | 529  | Tyr oksidaza         |
| ud11(ugt1)   | Hyperbilirubinemia; Crigler-Najjar syndrome      | 1 ? | 17  | 533  | Bilirubin transf     |
| cav3(cav3)   | Muscular dystrophy; Rippling muscle disease      | 1/2 | 7   | 151  | Adhezija             |