SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 3258

SASTAVLJANJE OPTIČKIH MAPA: MODUL ZA KOREKCIJU GRAFA

Luka Šterbić

Zagreb, lipanj 2013.

Sadržaj

Popis slika

1. Uvod

Temeljito poznavanje sekvence genoma nekog organizma omogućuje detaljan uvid u svojstva svih kromosomskih regija toga genoma. Ako sekvenca nije poznata, kao kod velikog broja eukariota, moguće je konstruirati mape prepoznatljivih genomskih elemenata. Mapa genoma prikazuje s određenom rezolucijom strukturnu organizaciju genoma, a sastoji se od oznaka, markera, za koje je poznata lokacija unutar genoma.

Iako je 99% ljudske sekvence DNA isto, varijacije u sekvenci DNA mogu imati snažan utjecaj na način na koji ljudi reagiraju na bolesti, okolišne faktore, viruse, bakterije, toksine, lijekove, itd. U proteklih nekoliko godina, medicinska istraživanja su se u velikoj mjeri fokusirala na pronalaženje genetskih uzroka bolesti. Pronađene su patologije asocirane jedinstvenom genetskom uzroku, ali broj otkrivenih poligenskih bolesti je velik i neprestano raste. Posljedica toga je promjena subjekta genetičkih istraživanja od jednoga gena na cjelokupni genom.

Mape genoma postaju ključni faktor za razvoj novih farmaceutskih proizvoda i kliničkih dijagnoza. Znanstvenici vjeruju da će mape pomoći pri identifikaciji mnogobrojnih gena koji se asociraju uz kompleksne bolesti kao što su rak, diabetes i kardiovaskularne bolesti, a kao budući cilj postavljaju individualnu kliničku analizu genoma.

Optičko mapiranje je metoda analize cjelokupnoga genoma, koju su predložili David C. Schwartz i drugi 1995. godine [1]. Temelji se na generiranju uređenih restrikcijskih mapa za cijele genome korištenjem metode iz molekularne biologije i mikrofluidne tehnike koje su proizašle iz poluvodičke industrije. Tako generirane restrikcijske mape mogu otkriti insercije, delecije, inverzije i ponavljanje genetskog materijala, te služe za uspostavu korelacije između genotipa i fenotipa u kliničkoj medicini.

Nedavna istraživanja su pokazala da postoji jaka asocijacija između strukturnih aberacija i pojave raka. Trenutne tehnologije strukturalne analize genoma spore su i skupe, a uređene restrikcijske mape otkrivaju strukturne značajke genoma u rezoluciji koju premašuje samo poznavanje kompletne sekvence DNA. Fokus ovog rada je modul za korekciju grafa poravnanja koji je sastavni dio algoritma *de novo* asembliranja.

2. Mapiranje genoma

Cilj mapiranja genoma je mapa koja prikazuje strukturnu organizaciju promatranoga genoma. Kao kod drugih vrsta mapa, mapa genoma mora prikazivati poziciju određenih značajki, markera. U geografskoj mapi kao markeri se mogu koristiti rijeke, planine, gradovi, itd. Genomska mapa sastoji se od skupa markera čija je lokacija unutar genoma poznata, a marker je definiran kao bilo koji strukturni element genoma koji se jednostavno identificira i kojem je pridijeljena specifična pozicija unutar genoma.

Prema konvenciji, razlikuju se dvije vrste mapiranja genoma:

- genetsko mapiranje,
- fizičko mapiranje.

Genetsko mapiranje oslanja se na genetske tehnika za konstruiranje mapa koje prikazuju poziciju pojedinih gena i druga obilježja sekvence genoma. Fizičko mapiranje koristi metode iz molekularne biologije za direktnu analizu molekule DNA kako bi se direktno ustanovila obilježja sekvence. U ovom radu fokus je na optičkom mapiranju restrikcijskih lokusa koje spada pod fizičko mapiranje.

Prve genetske mape, koje su se pojavile početkom dvadesetoga stoljeća, za jednostavne eukariote poput vinske mušice (*lat. Drosophila melanogaster*) koristile su gene kao markere. U to vrijeme geni su razmatrani kao apstraktni entiteti zaslužni za prijenos svojstva kroz generacije i bili su pogodni kao markeri jer je njihovo očitanje u fenotipu bilo jednostavno. U slučaju vinske mušice, mape su prikazivale poziciju gena za boju očiju, boju tijela i oblik krila, čiji je efekt na fenotipu bio vidljiv golim okom ili mikroskopom.

Da bi neka karakteristika bila pogodna za genetsko istraživanje, moraju postojati barem dva oblika fenotipskog očitanja te karakteristike, npr. bijela i crvena boja očiju kod vinske mušice (Slika 1.). Preko 50 gena genoma vinske mušice bilo je mapirano do 1922. godine, ali njih devet utjecalo je na boju očiju.

Slika 1. Fenotipske razlike u boji očiju vinske mušice [5]

S vremenom se ovakav pristup pokazao lošim zbog limitiranog broja vizualno različitih fenotipa i zbog utjecaja više gena na jedno svojstvo fenotipa te se prešlo na biokemijsko razlikovanje fenotipa, npr. AB0 sustav krvnih grupa ili analiza proteina u krvnom serumu kod čovjeka. Velika prednost ovakvih markera je što većina bitnih gena ima veliki broj alela. Kod eksperimentalnih organizama mogu se promatrati efekti selektivnog razmnožavanja, dok se podatci o heritabilnosti ljudskih gena dobivaju analizom fenotipova članova jedne obitelji. Ako svi članovi promatrane obitelji imaju isti alel za određeni gen, ne mogu se iz toga izvesti nikakve korisne informacije. Zbog toga je potrebno pretpostaviti slučajno razmnožavanje (*eng. random mating*) kako bi se relevantne kombinacije alela pojavile unutar jedne populacije, što je olakšano ako za promatrani gen postoje više od dva alela.

Geni su prikladni za markere, ali nisu idealni. Do problema dolazi kod većih genoma, kao onih kralježnjaka ili biljaka, jer mapa bazirana isključivo na genskim markerima nije jako detaljna. Čak i kad bi svaki gen bio mapiran, razina detalja bi ostala poprilično niska jer su kod većine eukariotskih genoma dva susjedna gena udaljena s velikim rupama između njih, ljudski genom ima oko tri milijarde nukleotida, ali oko trideset tisuća gena. Treba uzeti u obzir i da samo dio gena postoji u aleličnoj formi koju se može povoljno prepoznati. Genske mape nisu jako razumljive za više organizme te se dolazi do potrebe za drukčijim tipom markera.

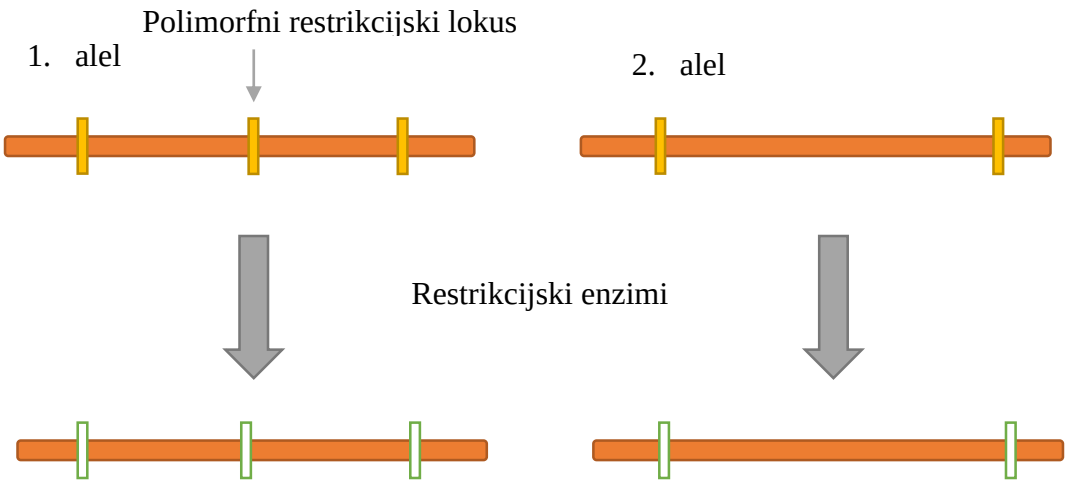
Mapirana svojstva koja nisu geni nazivaju se DNA markeri. Kao kod genskih markera, DNA marker mora imati barem dva različita alela da bi bio upotrebljiv za mapiranje. Postoje tri vrste sekvencijskih svojstva koje zadovoljavaju danu premisu:

- RFLP (*eng. restriction fragment length polymorphism*),
- SSLP (*eng. simple sequence length polymorphism*),
- SNP (*eng. single nucleotide polymorphism*).

2.1. RFLP markeri

RFLP markeri (*eng. restriction fragment length polymorphism*) su povijesno prvi DNA markeri koji su proučavani. Polimorfizam se u duljini restrikcijskog fragmenta veže uz pojam restrikcijskih enzima (detaljnije u nastavku rada, vidi Restrikcijski enzimi) koji režu molekulu DNA prilikom prepoznavanja specifične sekvence. Takav postupak bi trebao biti primjenjiv više puta s jednakim rezultirajućim skupom fragmenata jer je specifična sekvenca za određeni enzim ista, ali to nije uvijek slučaj jer su neki restrikcijski lokusi polimorfni. Polimorfni restrikcijski lokusi postoje kao dva alela. Jedan alel sadrži ispravnu sekvencu za restrikcijski lokus te će ona biti prepoznata od strane primijenjenog enzima koji će odrezati molekulu DNA na tom lokusu. Drugi alel ima takvu mutaciju sekvence zbog koje restrikcijski enzim više ne prepoznaje restrikcijski lokus i ne reže molekulu DNA. Rezultat alternacije sekvence u drugom alelu je povezanost dvaju susjednih restrikcijskih fragmenata nakon primjene enzima, što dovodi do polimorfizma u duljini fragmenata te se takva pojava karakterizira kao RFLP marker. Trenutno je poznato više od sto RFLP-ova u ljudskom genomu, ali za svaki RFLP mogu postojati samo dva alela.

Vrijednost RFLP-ova u istraživanju ljudskoga genoma je donekle limitirana zbog velike vjerojatnosti da članovi jedne promatrane obitelji ne prikazuju varijabilnost za promatrani RFLP.

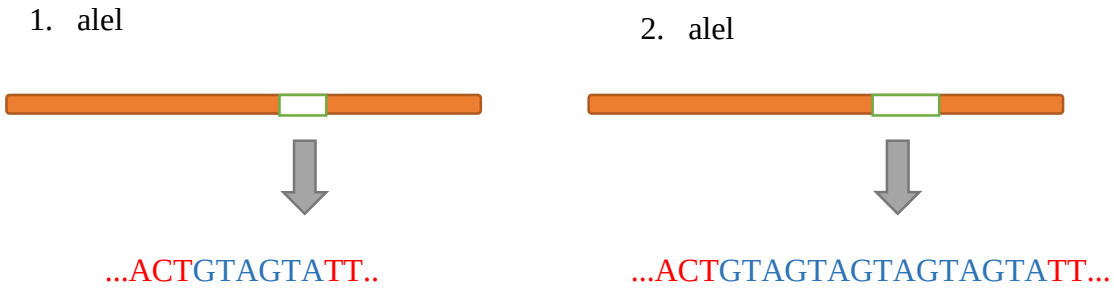


Slika 2. Djelovanje restrikcijskog enzima na polimorfni restrikcijski lokus

2.2. SSLP markeri

SSLP (*eng. simple sequence length polymorphism*) ili polimorfizam u duljini jednostavne sekvence je polje ponavljajuće sekvence koje pokazuje varijaciju u duljini ponavljanja, tj. različiti aleli sadrže različite brojeve ponavljanja sekvence. Razlikuju se dvije vrste SSLP-a:

- Minisateliti (*VNTR, eng. variable number of tandem repeats*), duljina ponavljajuće sekvence do 25 baza;
- Mikrosateliti (*STR, eng. simple tandem repeats*), kraća duljina ponavljajuće sekvence, najčešće dinukleotid ili trinukleotid.



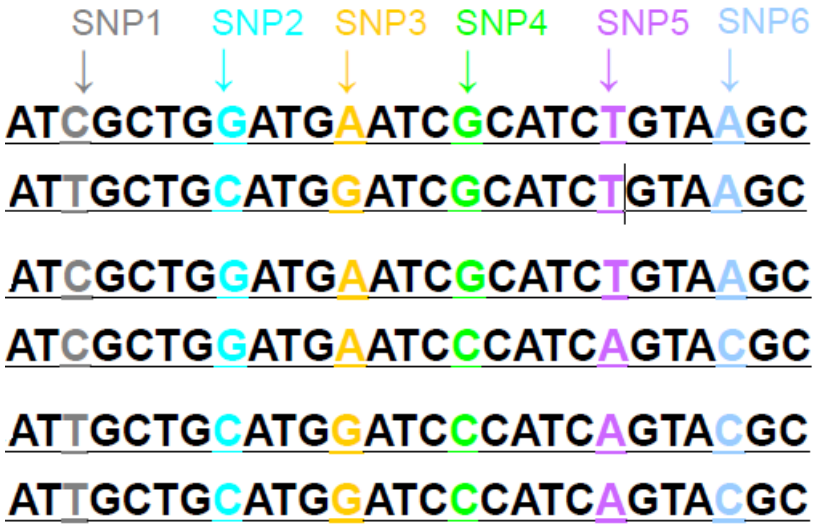
Slika 3. Primjer mikrosatelita s ponavljajućim tripletom „GTA“

Mikrosateliti su prikladniji markeri nego mikrosateliti. Mikrosateliti su ravnomjerno raspoređeni po cijelom genomu, dok minisateliti nisu. Minisateliti se češće nalaze u telomerskoj regiji pri krajevima kromosoma te kada bi ih se koristilo kao markere sredina kromosoma bi ostala slabo pokrivena. Najbrži način određivanja duljine ponavljajuće sekvence je primjena tehnike lančane reakcije polimerazom (*PCR, eng. polymerase chain reaction*). Tehnika PCR je brža i preciznija kod sekvenci do 300 baza duljine, što nije slučaj kod minisatelita zbog relativno velike duljine ponavljajuće sekvence. Mikrosateliti se tipično sastoje od 10 – 30 ponavljanja sekvence koja nije dulja od 4 baza, te su zato prikladniji za PCR. Ljudski genom sadrži oko $6.5 * 10^5$ mikrosatelita.

2.3. SNP markeri

SNP (*eng. single nucleotide polymorphism*), polimorfizam jednog nukleotida, često nazivan i „snip“, je pozicija u genomu na kojoj među jedinkama promatrane populacije postoji varijabilnost nukleotida. U svakom genomu postoji velik broj SNP-ova, te da bi se neka alteracija nukleotida smatrala SNP-om mora se pojavljivati u barem 1% populacije. Uzrok

RFLP-a je SNP u ponavljajućoj sekvenci, međutim mali broj SNP-ova stvara RFLP-ove jer sekvencu u kojoj leži SNP ne prepoznaje ni jedan restrikcijski enzim.



Slika 4. Primjer polimorfizma jednog nukleotida (SNPs) [5]

U ljudskom genomu je poznato više od 1.8 milijuna SNP-ova. Zaslužni su za 90% varijabilnosti ljudskog genoma i pojavljuju se svakih 100 – 300 baza uzduž 3 milijarde baza genoma. Na uzorku od tri SNP-a dva podrazumijevaju zamjenu citozina (C) s timinom (T). SNP-ovi se mogu pojavljivati u kodirajućim (genima) i nekodirajućim regijama genoma. Mnogo njih nema utjecaja na funkcije stanice, pogodnost organizma ili fenotip te se po teoriji o neutralnoj molekularnoj evoluciji smatraju neutralnima (teorija o neutralnoj molekularnoj evoluciji, Motoo Kimura, 1968; [5]).

Svaki SNP teoretski može postojati u četiri forme A, C, T, i G, tj. može imati četiri alela. Međutim, ustanovljeno je da to nije slučaj, već da većina SNP-ova postoji u samo dvije forme. SNP kao marker ima slične nedostatke kao RFLP maker u mapiranju ljudskog genoma, postoji velika vjerojatnost da u promatranoj obitelji ne postoji varijabilnost na određenom SNP-u.

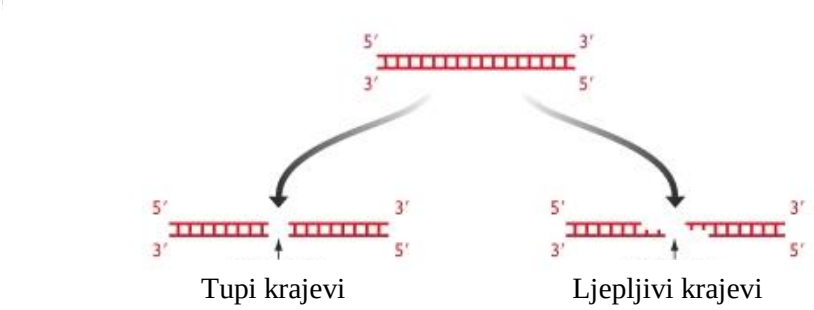
Velika prednost SNP markera je njihov veliki broj i raspršenost po genomu te mogućnost ispitivanja metodama koje ne koriste elektroforezu gelom koja se pokazala teškom za automatiziranje. Detekcija SNP-a je brža jer je osnovana na oligonukleotidnoj hibridizacijskoj analizi. Uz navedena svojstva, SNP-ovi su i evolucijski stabilni, ne dolazi do velikih promjena iz generacije u generaciju, što olakšava njihovo praćenje u populacijskim istraživanjima.

2. Restrikcijsko mapiranje

2.1. Restrikcijski enzimi

Mapiranje ljudskoga genoma znatno je ovisilo o mapiranju restrikcijskih enzima. Restrikcijski enzimi ili restrikcijske endonukleaze su specijalizirani proteini koji imaju sposobnost prekidanja fosfatne veze među bazama sekvence DNA na mjestima ili u neposrednoj blizini mjesta gdje prepoznaju specifični uzorak sekvence.

Otkriveni su sedamdesetih godina prošloga stoljeća te je 1978. godine dodijeljena Nobelova nagrada za fiziologiju i medicinu Werneru Arberu, Danielu Nathansu i Hamiltonu O. Smithu za njihov rad na otkriću i kategorizaciji restrikcijskih enzima. Restrikcijske endonukleaze su enzimi koje bakterija koristi za razgradnju strane DNA (virusna DNA). Enzim prepoznaje određenu nukleotidnu sekvencu (restrikcijsko mjesto ili lokus) od 4 do 8 nukleotidnih parova te na tome mjestu cijepa dvolančanu DNA. Takvo cijepanje može ostaviti tupe ili ljepljive krajeve molekule.

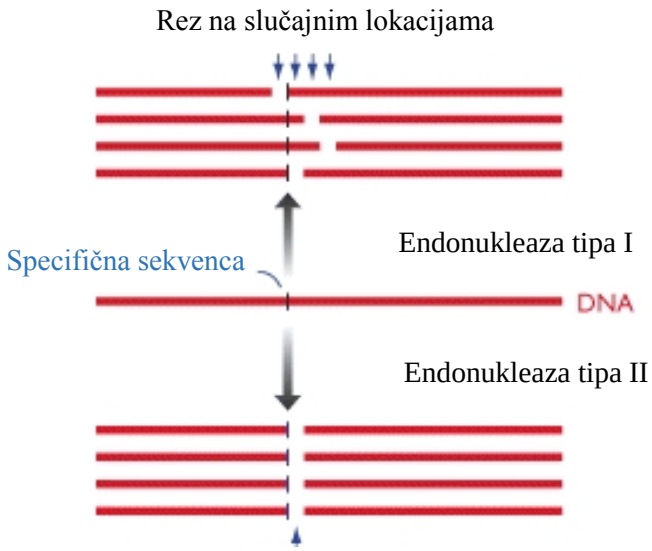


Slika 5. Primjer tupih i ljepljivih krajeva na molekuli DNA [3]

Danas je poznato preko 3000 restrikcijskih enzima, a njih 600 je dostupno komercijalno; njihovo otkriće je omogućilo razvoj tehnologije rekombinantne NDA koja ima veliku primjenu u medicini, npr. u proizvodnji insulina u velikim količinama za dijabetičare koristeći bakteriju Escherichia coli.

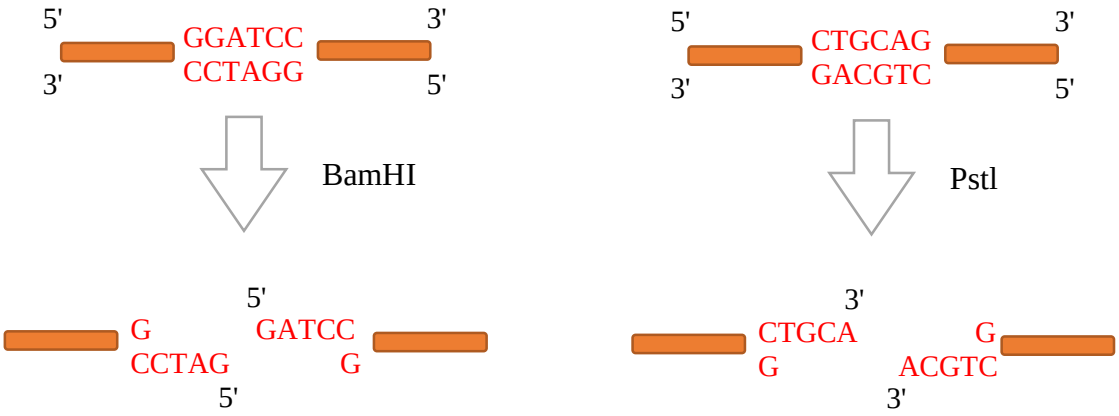
Restrikcijski enzimi su kategorizirani u četiri tipa: tip I, tip II, tip III i tip IV. Tipovi se međusobno razlikuju po enzimatskoj strukturi, vrsti sekvence koju prepoznaju i po mjestu cijepanja molekule DNA, a prilikom rezanja molekule svi restrikcijski enzimi stvaraju dva reza, jedan po svakom lancu spirale DNA.

Restrikcijski enzimi tipa I su prvi koji su otkriveni u E. coli i režu DNA na slučajnoj lokaciji u blizini prepoznatog uzorka sekvence. Udaljenost prepoznatog uzorka i lokacije reza iznosi barem 1000 baza. Restrikcijsko mjesto je asimetrično i sastoji se od dva specifična dijela. Prvi specifični dio, dužine 3 – 4 nukleotida, i drugi specifični dio, dužine 4 – 5 nukleotida, odvojeni su nespecifičnim dijelom dužine 6 – 8 nukleotida.



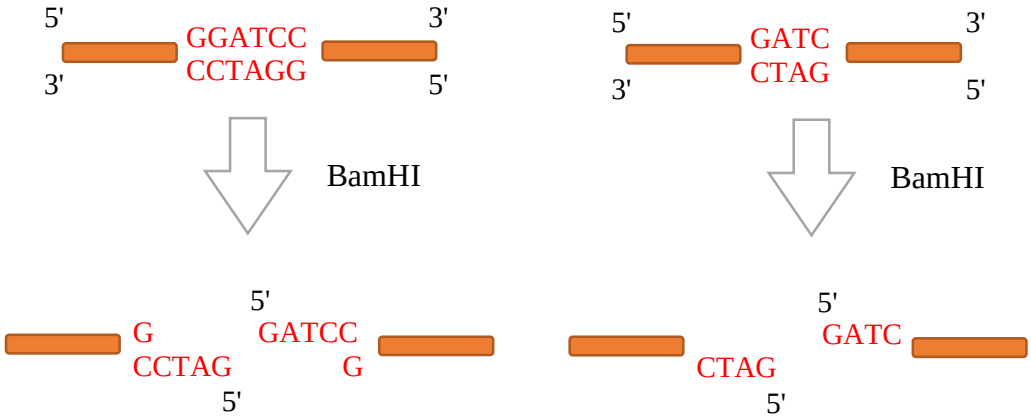
Slika 6. Razlika između endonukleaza tipa I i II [3]

Restrikcijski enzimi tipa II prepoznaju točno određena restrikcijska mjesta i na tim mjestima rade rez. Njihovi specifični uzorci su nepodijeljeni, 4 – 8 nukleotida dugi i često su palindromi. Restrikcijske endonukleaze ovog tipa nakon cijepanja ostavljaju ljepljive krajeve (npr. BamHI) ili tupe krajeve. Ovaj tip enzima ima danas najširu primjenu u industriji i istraživanju.



Slika 7. Primjer 5' i 3' overhanga

Različiti enzimi mogu proizvesti iste ljepljive krajeve, npr. BamHI i Sau3AI:



Slika 8. Isti ljepljivi krajevi kod različitih restrikcijskih enzima

Restrikcijski enzimi tipa III režu lanac DNA na lokaciji koja je blizu mjesta gdje je pronađena specifična sekvenca, obično 20 do 30 baza udaljenosti. Kod prokariota, ovaj tip enzima je dio mehanizma koji štiti jedinku od stranoga genetskog materijala. Specifična sekvenca koju enzim prepoznaje nije palindrom kao kod tipa II. Enzim prepoznaje kratke asimetrične sekvence (5 – 6 nukleotida) i reže 25 – 27 baza nizvodno i stvara kratka jednolančana izbočenja (5' overhang). Da bi došlo do restrikcije, potrebna su dva inverzno orijentirana nemetilirana restrikcijska mjesta, tj. dovoljno je da jedan lanac DNA bude metiliran pa da ne dođe do rezanja.

Posljednji tip restrikcijskih enzima, tip IV, je tip endonukleaza koji cilja modificiranu DNA (npr. metilirana). Nije prirodnog, već umjetnog podrijetla. Specifične sekvence takvih enzima mogu biti iznimno duge, do 36 nukleotida. Najkorištenije takve nukleaze su cink finger nukleaze (*eng. ZFNs*) koje se koriste u kloniranju i u genetskom inženjerstvu općenito.

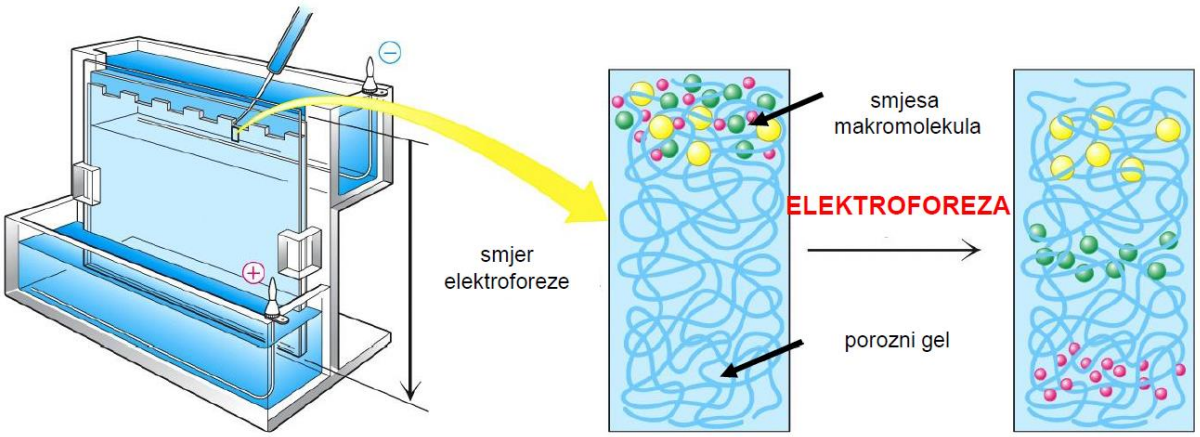
Mapiranje lokacija unutar genoma nad kojima djeluju opisani enzimi proizvodi restrikcijsku mapu koja prikazuje prostorne informacije za specifične genetske lokuse. Ovom je metodom uspješno mapirano nekoliko bakterija poput E. Coli, S. cerevisiae i C. elegans, ali se problem pojavio prilikom mapiranja viših organizama.

Presudna tehnologija za ovakav tip mapa je gel elektroforeza koja služi za određivanje dužina segmenata molekula DNA. Dugo se zbog manjka softwarea za automatiziranu gel elektroforezu, niske propusnosti cjelokupne metode i potrebne količina ručnog rada kočio razvoj restrikcijskog mapiranja kao metode analize cjelokupnoga genoma.

2.2. Gel elektroforeza

Elektroforeza je metoda razdvajanja i analize makromolekula (DNA, RNA i proteini) u električnom polju na temelju razlike u naboju i veličini molekula pri određenom pH. Izvodi se na inertnim podlogama kao što su celuloza-acetat ili agar gel, odakle naziv gel elektroforeza.

Kada se kroz emulziju propusti istosmjerna struja, čestice se gibaju prema katodi ili anodi određenom brzinom. Brzina gibanja molekula u električnom polju ovisi o netto naboju, masi i obliku molekula te o jakosti električnog polja i svojstvima medija kroz koji se gibaju.



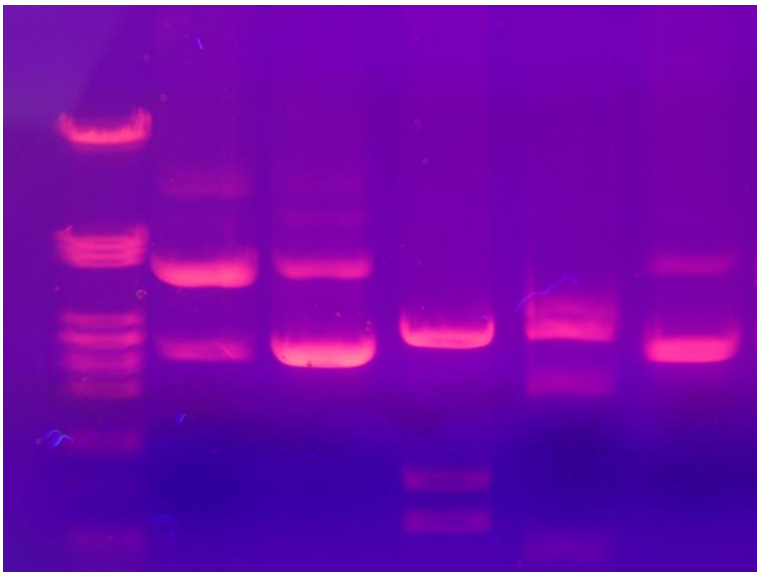
Slika 9. Proces gel elektroforeze

Elektroforeza molekule DNA se najčešće izvodi u poroznim gelovima (gelovi poliakrilamida, gelovi agaroze) koji služe kao mehanički nosači mobilne faze te djeluju kao "molekularna sita", tj. što su molekule veće to sporije putuju kroz gel, dok su manje molekule brže. Osim toga, gel služi i za ublažavanje termičkih efekata električne struje na analizirane molekule.

Nakon završetka elektroforeze, makromolekule se u gelu mogu obojati kako bi postale vidljive. DNA se može vizualizirati npr. primjenom etidij bromida koji je fluorescentan pod utjecajem ultraljubičastog zračenja, dok se proteini mogu vizualizirati bojanjem srebrom ili bojanjem Coomassie bojom (*eng. Coomassie Brilliant Blue dye*). Ako su molekule koje se separiraju radioaktivne ili sadrže radioaktivne elemente, umjesto bojanja se za detekciju molekula može snimiti autoradiogram gela. Autoradiogram je slika na filmu koji reagira na X

zrake koje nastaju zbog emisija elementarnih čestica koje se javljaju prilikom radioaktivnog raspada (gamma zračenja, alfa i beta čestice).

Najčešći oblik bojanja za vizualizaciju DNA i RNA u agaroznim gelovima je etidij bromid (EtBr). Ako se gel tretiran etidij bromidom osvijetli UV svjetlom nakon što elektroforeza završi, svaki će dio koji sadrži barem 20 ng DNA biti jasno vidljiv. Poznato je da EtBr ima mutagena svojstva, što je dovelo do razvoja sigurnijih varijanti kao što je GelRed.



Slika 10. Gel elektroforeza molekule DNA uz primjenu etidij bromida i UV zračenje za vizualizaciju [7]

3. Optičko mapiranje

3.1. Razvoj metoda optičkog mapiranja

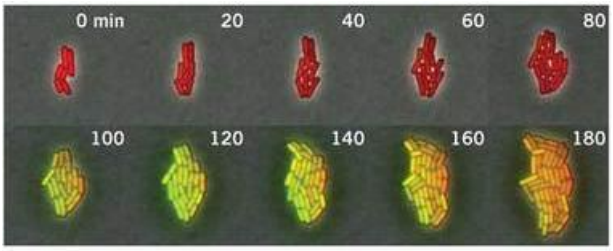
Optičko mapiranje je metoda analize cjelokupnoga genoma, koju su predložili David C. Schwartz i drugi 1995. godine. Temelji se na generiranju uređenih restrikcijskih mapa za cijele genome koristeći metode iz molekularne biologije i mikrofluidne tehnike koje su proizišle iz poluvodičke industrije.

Ova se metoda bitno razlikuje od klasičnoga restrikcijskog mapiranja gel elektroforezom u smislu propusnosti. Jedna od razlika je u molekulama koje se koriste u procesu. Optičko mapiranje analizira restrikcijske fragmente jedne molekule DNA, dok mapiranje elektroforezom koristi veliki skup fragmenata iz više molekula DNA.

Optičko mapiranje može biti kompletno automatizirano zahvaljujući tehnologijama iz područja računalnog vida i obrade slike, za razliku od klasičnoga restrikcijskog mapiranja kod kojega je potrebna velika količina ručnog rada. Ova tehnologija je sposobna proizvesti restrikcijske mape visoke rezolucije bez prethodnog poznavanja sekvence promatrane molekule DNA.

Tako generirane restrikcijske mape mogu otkriti insercije, delecije, inverzije i ponavljanje genetskog materijala te služe za uspostavu korelacije između genotipa i fenotipa u kliničkoj medicini.

U originalnom postupku optičkog mapiranja, fluorescentno označene molekule DNA su rastegnute na geloznoj bazi agara u kojoj su prethodno ugrađeni restrikcijski enzimi. Ioni magnezija Mg^{2+} aktiviraju restrikcijske enzime i pokreću proces probave. Mjesta na kojima su restrikcijske endonukleaze obavile probavu i izrezale molekulu postaju vidljiva kao praznine nakon što se fragmenti DNA povuku zbog elasticiteta molekule. Kako molekule nisu čvrsto fiksirane u agaru, za analizu rezultata nije dovoljno pribaviti jednu sliku, već se koristi vremenski kontinuirana fluorescentna mikroskopija.



Slika 11. Primjer vremenski kontinuirane fluorescentne mikroskopije [8]

Veliko ograničenje ovakvog postupka je u metodi pribavljanja podataka. Problemi proizlaze iz činjenice da molekule DNA unutar agara nemaju organiziranu strukturu te je lokacija takvih fragmenata slučajna, što znatno otežava pribavljanje kvalitetne slike zbog kontinuiranog snimanja.

Druga generacija optičkog mapiranja zamijenila je agarsku podlogu staklenom podlogom tretiranom polilisinom. Pozitivno nabijeno staklo elektrostatski interagira s negativno nabijenim molekulama DNA i pričvršćuje ih za podlogu. Takvo pričvršćivanje molekula izbacilo je potrebu za kontinuiranom fluorescentnom slikom jer ovim postupkom molekule ostaju u fokusu mikroskopa.

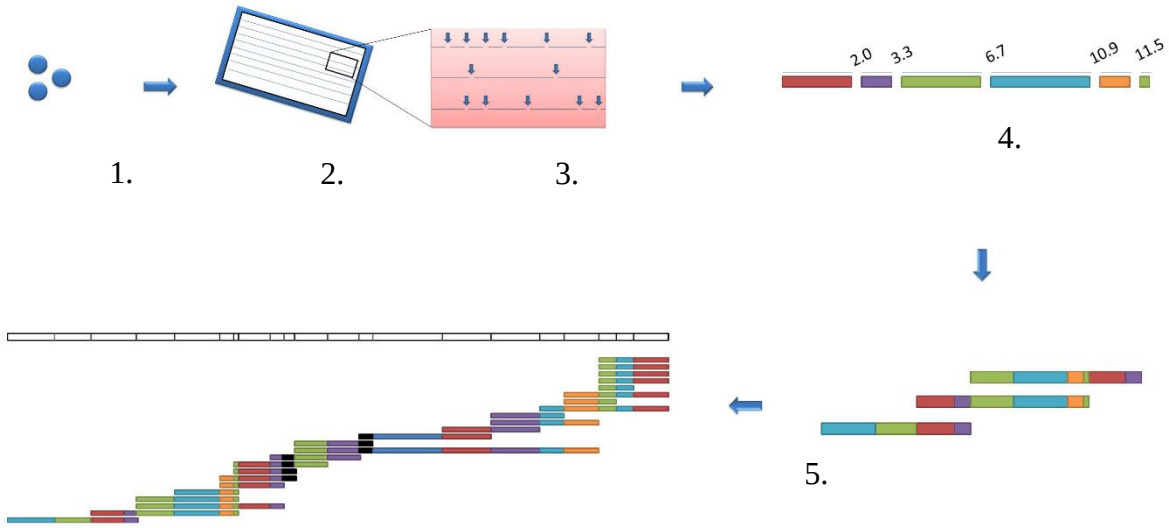
Ova generacija optičkog mapiranja temelji se na konstrukciji mape iz velikog broja malih fragmenata koji se preklapaju. Analiza pribavljene slike identificira točan broj fragmenata koji su nastali djelovanjem endonukleaza i gradi histograme u ovisnosti o veličinama fragmenata. Analizom histograma ustanovljuju se srednje veličine fragmenata iz kojih je jednostavno izgraditi restrikcijske mape jer je poredak restrikcijskih fragmenata optičkim mapiranjem očuvan. Rezolucija ove metode proteže se od 800 bp do ~30 kbp.

Moderna je metoda optičkog mapiranja poboljšanje prethodno opisanih metoda u smislu automatizacije postupka. Svaki korak od pribavljanja i purifikacije DNA iz stanica do pribavljanja i analize slike mikroskopom može se automatizirati.

Moderno optičko mapiranje provodi se u nekoliko koraka (vidi Slika 12. Tijek operacija prilikom postupka optičkom mapiranja):

1. Iz liziranih stanica dobiva se genomska DNA koja se sjecka na manje komadiće koji čine biblioteku molekula za optičko mapiranje.
2. Jedna, negativno nabijena, rastegnuta molekula DNA postavlja se i pričvršćuje na staklo s pozitivnim nabojem nad kojim je postavljen mikroskop.

3. Dodaju se restrikcijski enzimi koji režu rastegnutu molekulu na mjestima gdje prepoznaju specifičan uzorak sekvence. Fragmenti ostaju pričvršćeni za podlogu i povlače se prema središtu zbog elasticiteta molekule DNA. Nastaju rupe u linearnoj strukturi molekule koje se mogu jednostavno identificirati mikroskopom.
4. Zbog fluorescentnog bojanja, fragmenti DNA će biti vidljivi mikroskopom pod utjecajem ultraljubičastog zračenja (fluorescentna mikroskopija), a veličina fragmenata određuje se na bazi intenziteta fluorescencije. Što je segment duži, to će zračenje biti jače i obratno. Rezultat ovog koraka je optička mapa jedne molekule.
5. Mape pojedinih molekula međusobno se kombiniraju i stvara se suglasna mapa cjelokupnoga genoma.



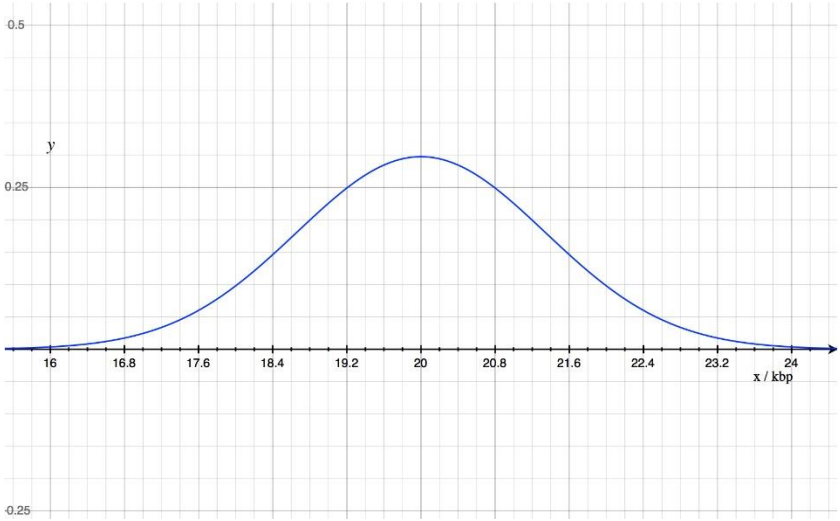
Slika 12. Tijek operacija prilikom postupka optičkom mapiranja [10]

3.2. Analiza podataka u optičkom mapiranju

Količina podataka koja se generira optičkim mapiranjem iznimno je velika. Analizirati takve podatke s ciljem konstruiranja što preciznije globalne mape izrazito je kompleksno zbog imperfektnosti ulaznih podataka. Procesom mapiranja nastaju pogreške koje je potrebno identificirati i, ako je to moguće, ispraviti.

U postupku se najčešće koriste restrikcijski enzimi sa specifičnom sekvencom dužine 6 bp. Svaka mapa predstavljana je kao polje veličina fragmenata poredanih kako su očitani sa slike. Pojedine mape variraju po veličini od 350 kbp do 4 Mbp i sastoje se od tridesetak

fragmenata. Oko 20% restrikcija nije uspješno obavljeno i obično se pojavljuju tri lažna proreza po 1 Mbp sekvence. Većina fragmenata manjih od 500 bp nije reprezentirano u ulaznom skupu podataka, dok su fragmenti do 2 kbp slabo reprezentirani. Veličina restrikcijskih fragmenata X podliježe normalnoj distribuciji $X \approx N(Y, \sigma^2 \cdot Y)$ gdje je $\sigma^2 \approx 0.3$, a Y stvarna veličina



Slika 13. Distribucija očitanih veličina za fragment veličine 20 kbp

fragmenta. Npr. Za segment stvarne veličine 20 kbp, 80% mjerenja je unutar 3,3 kbp od te vrijednosti.

Nakon reagiranja enzima, na podlozi ostaju usmjerene i označene molekule DNA. Slike koje pribavlja automatizirani sustav prikazuju isjeckane molekule DNA na kojima su vidljive „rupe“ veličine ≈1 μm na mjestima gdje je došlo do reakcije. Relativna veličina fragmenata izračunava se na bazi intenziteta fluorescencije fragmenata i internoga standarda poznate dužine (λ DNA) koji je dodan podlozi prije snimanja slike. Proces rezultira skupom slika koje sadrže restrikcijske fragmente poznate dužine nalik na bar kodove.

Zbog krhkosti molekule DNA, teško je dobiti intaktnu molekulu na podlozi prije reakcije restrikcijskih enzima. Upravo se zbog toga optičko mapiranje temelji na velikom broju slučajno isjeckanih lanaca DNA, od kuda i naziv „Shotgun Optical Mapping“. Velikom redundantnošću nesavršenih ulaznih podataka nastoje se minimizirati pogreške koje nastaju zbog tehničke prirode postupka na način da svaka genomska regija bude višestruko pokrivena u uzorcima. U takve greške se ubrajaju:

- Lažna restrikcijska mjesta,
- Neefikasnost reakcije restrikcije (oko 70 – 90%),
- Slaba podatkovna pokrivenost fragmenata manjih od 2 kbp,

- Greške prilikom mjerenja veličine fragmenata,
- Kimerne ili umjetne mape.

Generiranje precizne mape genoma pospješuje se statističkom analizom više nesavršenih mapa i nedavnim napredovanjima u algoritmima koji se temelje na teoriji grafova. Pokazalo se da kombiniranjem rezultata više postupaka optičkog mapiranja u pravilu rezultira kvalitetnijim konačnim rezultatom.

Kod shotgun optičkog mapiranja pojedina mapa predstavlja slučajnu lokaciju unutar genoma, a ne cijelu molekulu. Zbog toga se nastoji postići da svaka analizirana regija bude pokrivena u 10 do 50 mapa. Mnogo je istraživanja napravljeno na temi rekonstrukcije globalne mape iz mapa malih molekula. U općenitom slučaju problem se predstavlja kao NP-težak, no pronađeni su algoritmi koji barataju polinomnim složenostima za niže eukariote poput bakterija. Nažalost, takvi algoritmi nisu primjenjivi prilikom shotgun optičkog mapiranja i kod viših organizama poput čovjeka zbog velike prostorne i vremenske složenosti.

Anton Valouev, David C. Schwartz, Shiguo Zhou i Michael S. Waterman predložili su novi algoritam ([1]), prvi takve vrste, za de novo sastavljanje optičke mape cijeloga genoma. Algoritam kao ulazne podatke prima optičke mape slučajno isjeckanih molekula DNA (shotgun optical mapping) i primjenjiv je na višim eukariotima poput čovjeka u prihvatljivom vremenu, uz pretpostavku izvođenja na računalnom klasteru.

Algoritam se sastoji od tri slijedna koraka i temelji se na modificiranom računalnom okviru preklapanje – tlocrt – suglasnost (*eng. overlap – layout – consensus framework*) koji se često koristi u sekvencijskim asemblerima. Postupak optičkog asembliranja kompleksniji je od sekvencijskoga zbog toga što restrikcijske mape jedinstvene molekule sadrže greške.

Prvi korak algoritma, preklapanje, zadužen je za uspostavu veze između ulaznoga skupa optičkih mapa. Korak tlocrta razvrstava veze među mapama na lokalne i globalne, a korak suglasnosti stvara konačnu globalnu restrikcijsku mapu. Povezanost mapa predstavlja se usmjerenim grafom i koristi sustav za korekciju pogrešaka baziran na udaljenosti. Skica konačne mape generira se iz usmjerenoga grafa sastavljanjem više pojedinih mapa. Zadnji korak prije generiranja konačne mape je proces rafiniranja rezultata koji može otkriti i ispraviti određene pogreške nastale tijekom asembliranja.

3.3. Algoritam sastavljanja optičkih mapa

Algoritam sastavljanja odvija se u sedam slijednih koraka:

1. Izračun preklapanja,
2. Konstrukcija grafa preklapanja,
3. Korekcije grafa,
4. Identifikacija otoka,
5. Konstrukcija susjedstva
6. Konstrukcija skice konačne mape
7. Uglađivanje konačne mape.

Prvi je korak, računanje poravnanja svih parova mapa u ulaznome skupu podataka, najzahtjevniji po pitanju računalnih resursa te se zato provodi samo jednom, a ostali koraci koriste već izračunate rezultate. Složenost ovoga koraka algoritma je kvadratna s obzirom na broj mapa. Ovakva struktura postupka omogućuje brzo ponovno sastavljanje mapa ako je došlo do promjene parametra asemblera.

Svakom poravnanju pridružuje se vrijednost koja služi kao oznaka značajnosti preklapanja, a samo kvalitetna poravnanja propuštaju se kroz sljedeće korake. Ako uzmemo u obzir da za ljudski genom postoji više od pola milijuna optičkih mapa, očita je potreba za paralelizacijom usporedbi. Uza znatne računalne resurse, prvi korak se može kompletirati u razumnom vremenu. Vremenska složenost poravnanja dviju mapa s m i n fragmenata je $O(mn)$, konkretno, oko 2,5 ms na prosječnom osobnom računalu. Prostorna složenost poravnanja nije velika i iznosi, kao i vremenska, $O(mn)$.

Kako bi se identificirala kvalitetna poravnanja, koristi se sistem ocjenjivanja koji uzima u obzir lažne procijepe, procijepe koji nedostaju te pogreške u određivanju dužine restrikcijskih fragmenata. Ocjena poravnanja nastoji nagraditi sa procijepe koji se podudaraju i penalizirati lažne ili nepostojeće proreze s λ. Svaki segment mape [i, k] sastoji se od lokacija reza i do k, a podudarajući par definiran je kao (i,j;k,l.). Globalna ocjena poravnanja dviju mapa s m i n procijepa definirana je kao skup uređenih podudarajućih parova. Lokacije proreza prikazane su sa q_x za lokaciju x unutar prve mape i r_y za lokaciju y unutar druge mape. Raskorak u udaljenosti podudarajućih parova uzet je u obzir korištenjem funkcije sličnosti duljina $l(a, b)$ gdje su a i b duljine promatranih mapa.

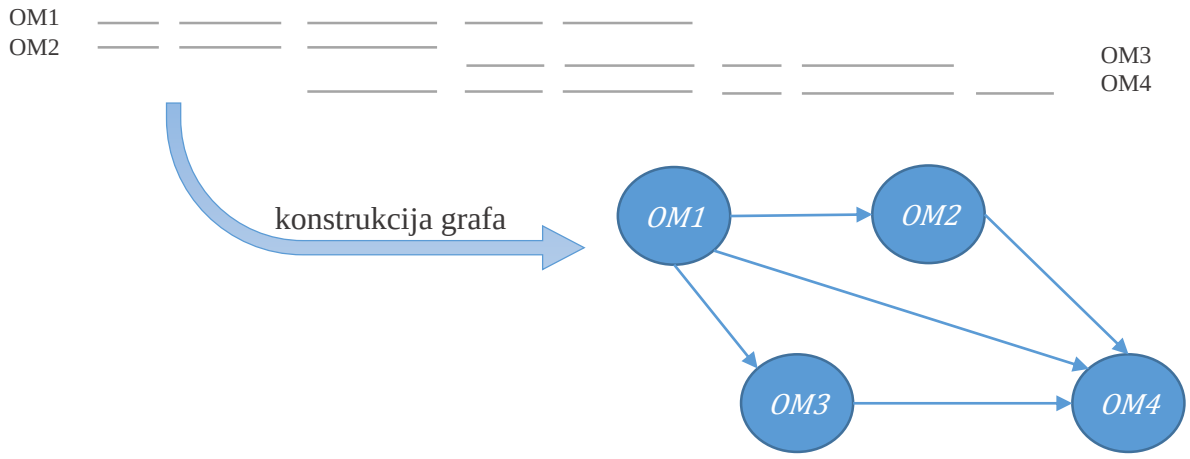
$$score(\Pi) = \sum_{t=1}^{d} \sigma(i_t, j_t; k_t, l_t) + l(q_{i1}, r_{i1}) + \sum_{t=2}^{d} l((q_{i_t} - q_{k_{t-1}}), (r_{j_t} - r_{l_{t-1}}))$$

$$+ l\left((q_m - q_{k_a}), (r_n - r_{l_d})\right) - \lambda[m + n - \sum_{t=1}^{d}(k_t - i_t + 1) - \sum_{t=1}^{d}(l_t - j_t + 1)]$$

$$\sigma(i_t, j_t; k_t, l_t) = v \cdot (broj\ podudarajućih\ parova\ u\ segmentu\ [i_t, j_t; k_t, l_t])$$

$$+ l\left((q_{k_t} - q_{i_t}), (r_{l_t} - r_{j_t})\right) - \lambda((k_t - i_t) + (l_t - j_t))$$

U drugom se koraku gradi usmjereni graf $G(V, E)$ koji predstavlja relacije preklapanja među pojedinim mapama. Optičke mape predstavljene su kao čvorovi grafa V, dok su preklapanja među mapama predstavljena kao usmjereni bridovi E. Ovakav oblik prikaza pogodan je zbog mogućnosti korištenja algoritama grafova kao što su pretraživanje u širinu, iterativno pretraživanje u dubinu i najteži put.

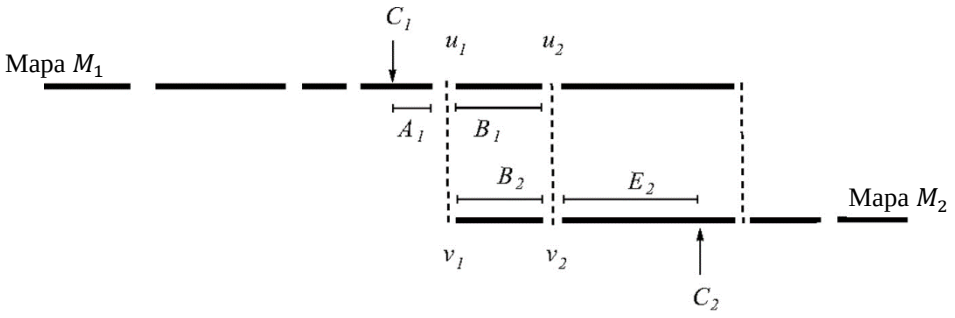


Slika 14. Primjer konstrukcije grafa iz skupa optičkih mapa

Ocjene poravnanja iz prvoga koraka koriste se kao prva razina filtriranja kvalitetnih preklapanja, a samo preklapanja koja su veća od određene granice (*eng. q-score*) uključuju se u graf. Lista poravnanja silazno se sortira tako da se kvalitetnija poravnanja ranije dodaju u graf. Takva metoda konstrukcije pospješuje minimizaciju broja lažnih bridova koji se dodaju u graf jer takvi bridovi tipično imaju niski q-score. U ovom se koraku vodi računa o mogućim greškama lažnih preklapanja s nekonzistentnom orijentacijom (detaljnije u 3.4).

Određivanje orijentacije bridova nije trivijalno jer ne postoji garancija da molekule DNA na podlozi imaju istu orijentaciju. Zbog toga svakoj mapi treba pridijeliti orijentaciju, normalnu ili obrnutu, s obzirom na stanje drugih mapa koje su već dodane u graf. Težina

bridova određuje se računanjem genomske udaljenosti preklapajućih mapa koja je definirana kao udaljenost centara mapa.



Slika 15. Računanje udaljenosti centra mape M_1 i M_2 [1]

Udaljenost centra C_1 i C_2 dana je izrazom $A_1 + \frac{(B_1+B_2)}{2} + E_2$, gdje u_1 predstavlja podudarajuću lokaciju najvećeg bloka preklapanja $(u_1, u_2; v_1, v_2)$ koja je najbliža C_1 i ne sadrži centre mapa C_1 i C_2. Predznak ovisi o orijentaciji pojedinih mapa. Tijekom konstrukcije grafa postavlja se smjer bridova tako da težina dotičnog brida bude pozitivna.

Treći korak, koji će biti detaljnije pojašnjen u idućem potpoglavlju, zadužen je za korekciju grafa. Bez ispravljanja grešaka nastalih u prethodnim koracima i u procesu mjerenja često nije moguće konstruirati preciznu globalnu restrikcijsku mapu. Greške u grafu mogu se pojaviti u obliku lažnih bridova (*eng. false edges*) i lažnih čvorova (*eng. spurious nodes*), a ukazuju na lažnu asocijaciju dviju regija genoma. Procedura korekcije grafa identificira i eliminira takve greške. Moguće greške jesu:

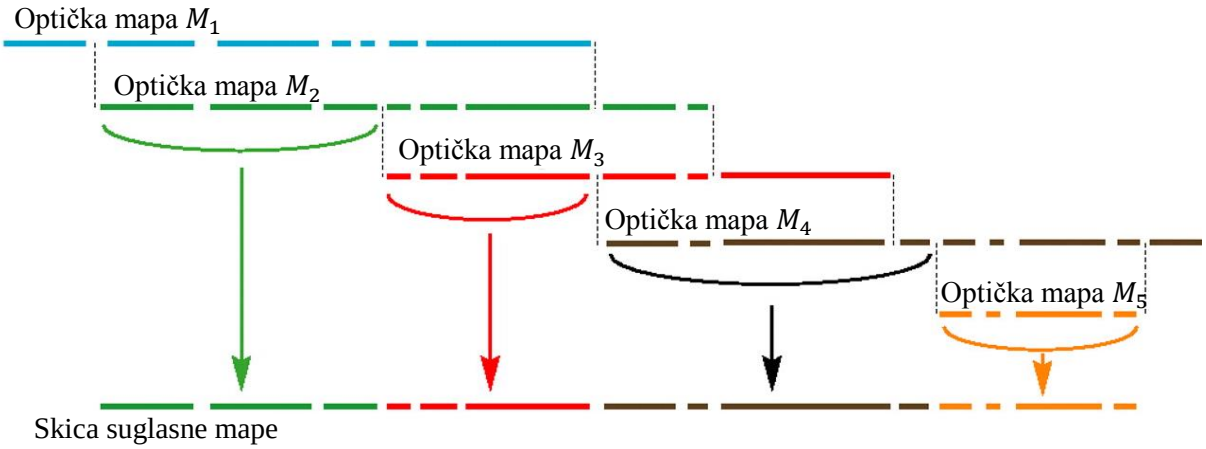
1. Lažni bridovi,
2. Lažna preklapanja s konzistentnom orijentacijom,
3. Lažna preklapanja s nekonzistentnom orijentacijom,
4. Kimerne mape.

Sljedeći korak je identifikacija „otoka“. Nakon korekcije, graf se raspada na nekoliko komponenta koje predstavljaju povezane regije genoma razapete povezanim optičkim mapama. Za svaku takvu komponentu, koje se naziva i otok, mora se izvaditi regija genoma koja pripada tom otoku, tj. susjedstvo. Svaki otok zasebno se obrađuje u daljnjim koracima kako bi se generirala suglasna mapa te regije.

U svakoj komponenti grafa, susjedstva mogu biti prikazana kao putevi koji spajaju izvore i ponore. Izvori su definirani kao čvorovi koji imaju samo izlazne bridove, dok su ponori čvorovi koji imaju samo ulazne bridove. Kako bi se producirala najopsežnija moguća mapa,

traži se najteži aciklički put u podgrafu. Takvim se postupkom maksimizira procijenjena genomska udaljenost koju put pokriva i to je zadaća petoga koraka, tj. konstrukcija susjedstva. Iz svakoga identificiranog izvora pokreće se pretraživanje u dubinu koje u svakom čvoru zapisuje duljinu najdužeg puta do toga čvora koja je trenutno poznata. Najteži put se određuje pronalaženjem čvora s najvećom zapisanom vrijednošću i puta koji završava u tom čvoru.

U šestom se koraku konstruira skica konačne suglasne mape iz puta koji je pronađen u prethodnom koraku tako da se spajaju pojedinačne mape na mjestima gdje se one preklapaju. Svaka tako konstruirana mapa predstavlja jedan otok, tj. jednu regiju genoma.



Slika 16. Konstrukcija skice suglasne mape [1]

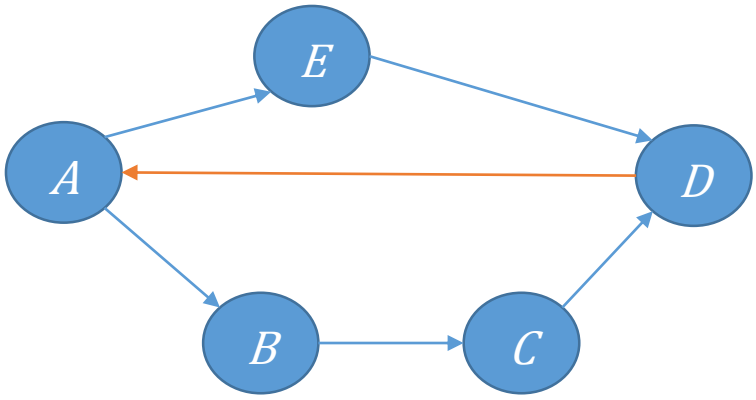
Posljednji, sedmi, korak algoritma zadužen je za pročišćivanje konačne mape. Konačna mapa nastaje konkatenacijom manjih optičkih mapa što uvodi mogućnost pogrešaka u obliku lažnih procijepa, izgubljenih procijepa i velike varijance u dužini fragmenata. Ako je skica konačne mape dovoljno detaljna, takve se pogreške mogu ispraviti kombiniranjem informacija iz velikog broja optičkih mapa. Tako nastala mapa je suglasna mapa analizirane regije genoma i isporučuje se kao rezultat procesa asembliranja.

3.4. Modul za korekciju grafa

Greške unutar grafa preklapanja mogu dovesti do asocijacija regija genoma koje u stvarnosti nisu povezane. Lažni bridovi predstavljaju nepostojeću povezanost genomskih regija, a lažni čvorovi predstavljaju tzv. kimerne mape koje nastaju zbog fizičkog preklapanja dvaju molekula DNA prilikom pribavljanja slike. Korekcija grafa mora biti provedena u smislu eliminacije lažnih bridova i čvorova.

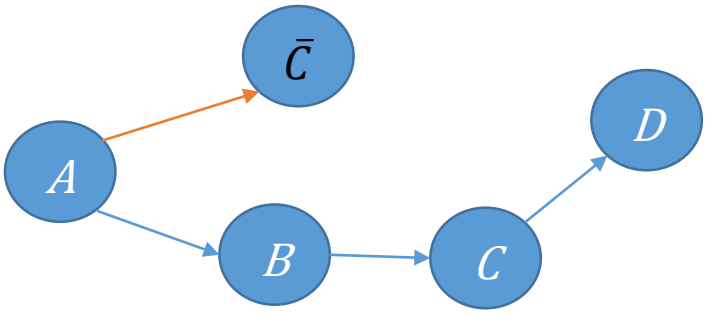
Lažni bridovi i kimerne mape mogu jako nalikovati područjima niske pokrivenosti i slabog preklapanja. Da bi dvije regije genoma ostale povezane nakon korekcije, algoritam mora pronaći više dokaza asocijacije. Svake dvije regije moraju biti povezane s barem dva puta kroz graf preklapanja koji nemaju zajedničke čvorove. Kako bi ovaj korak bio uspješan, pokrivenost analiziranoga genoma mora biti velika, ali se pokazalo da višestruka pokrivenost u konačnici dovodi do preciznijih asembliranih mapa.

Postoje tri vrste grešake bridova koje je potrebno eliminirati iz grafa preklapanja: lažni bridovi, lažna preklapanja s nekonzistentnom orijentacijom i lažna preklapanja s konzistentnom orijentacijom. Lažni bridovi povezuju dva čvora stvarajući ciklus unutar grafa. Takva struktura unutar linearnoga genoma nije moguća i stoga takve bridove treba izbrisati.



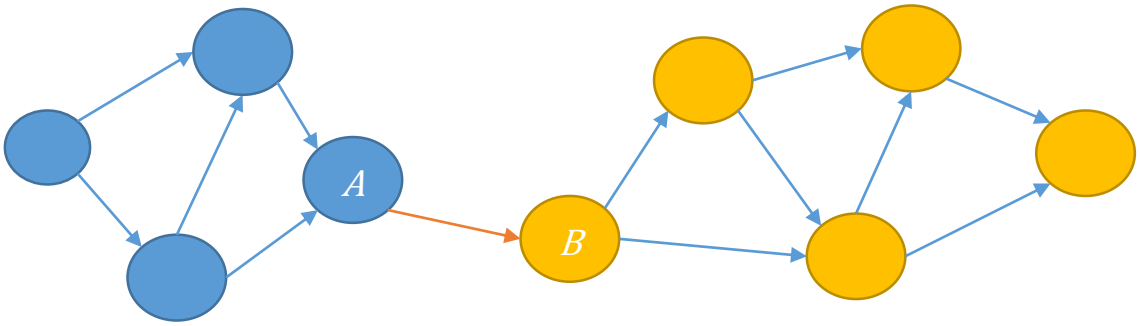
Slika 17. Primjer lažnog brida (D → A)

Bridovi koji predstavljaju preklapanje s nekonzistentnom orijentacijom stvaraju konflikt orijentacija u grafu, ali o ovoj vrsti greške bridova nije potrebno brinuti u ovom modulu jer se konzistentnost orijentacije provjerava prilikom konstrukcije grafa (2. korak algoritma). Tijekom konstrukcije grafa, svaki novi brid koji se dodaje prethodno postojećoj strukturi mora imati kompatibilnu orijentaciju, jer ako orijentacija nije kompatibilna takav se brid odmah odbacuje.



Slika 18. Primjer lažnog preklapanja s nekonzistentnom orijentacijom (A → $\bar{C}$)

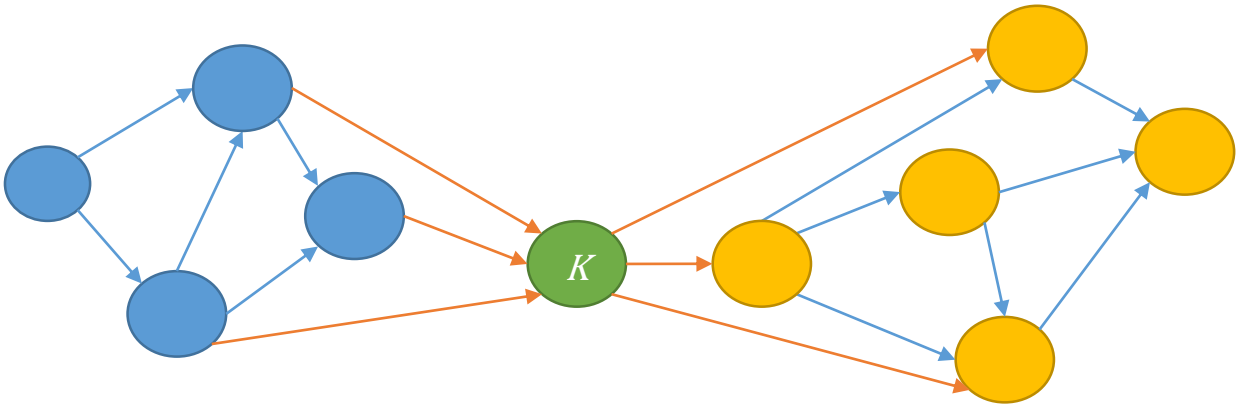
Lažna preklapanja s konzistentnom orijentacijom prikazana su bridovima koji spajaju nepovezane dijelove genoma. Za identificiranje takvih bridova potrebno je provesti ograničeno pretraživanje u dubinu za svaki čvor grafa N_i. Tijekom pretraživanja treba zapamtiti svaki čvor N_j za kojega postoje više nezavisnih puteva od N_i do N_j. Nezavisni putevi su oni putevi koji nemaju zajedničkih čvorova, osim početnog i krajnjeg čvora. Za svaki tako pronađeni put potrebno je izračunati udaljenost D_α koja maksimizira veličinu klastera puteva od jednoga čvora do drugog. Ako je pronađeno više puteva s normalno distribuiranim udaljenostima, svi bridovi tih puteva označavaju se kao potvrđeni. Svi bridovi koji u ovom postupku nisu potvrđeni brišu se kao i svi izolirani čvorovi.



Slika 19. Primjer lažnog preklapanja s konzistentnom orijentacijom (A → B)

Kimerne mape se tipično sastoje od dvije grupe čvorova koje su povezane preko samo jednoga čvora. U smislu fizičkog preklapanja kimerna mapa izgleda kao mapa čiji se lijevi dio preklapa s jednim skupom mapa, a desni dio s drugim skupom, dok ne postoji niti jedna druga mapa koja bi povezala mape iz lijevog i desnog skupa. Promatrana regija nije lokalno povezana niti sa jednom drugom regijom grafa, ako zanemarimo tu jednu točku, tj. potencijalne kimerne mape identificiramo kao lokalne artikulacijske čvorove.

Postupak identifikacije kimernih mapa koristi ograničeno pretraživanje u širinu s ciljem dokazivanja da bi brisanjem tekućeg čvora graf postao lokalno nepovezan. Za svaki čvor koji je susjedni potencijalnoj kimernoj mapi pokreće se pretraživanje tražeći sve neposredne susjede kimernog čvora tako s time da se ignoriraju svi putevi kroz potencijalni kimerni čvor. Čvor smatramo kimerni ako niti jedna instanca pretraživanja nije pronašla sve susjede čvora. Kad je kimerni čvor pronađen, brišu se čvor i svi pripadajući bridovi. Na kraju postupka se brišu svi izolirani čvorovi.

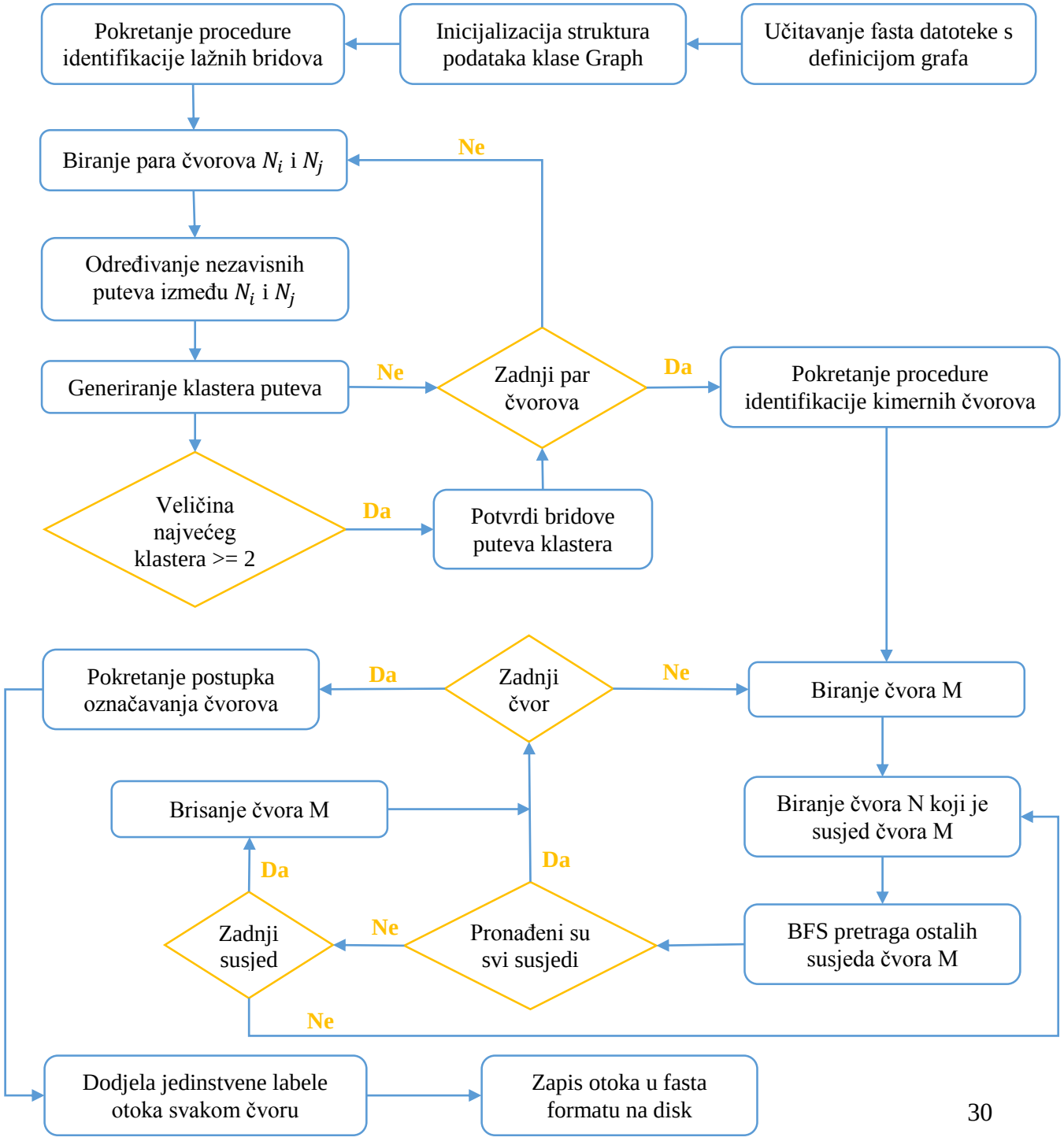


Slika 20. Primjer kimerne mape (K)

4. Implementacija

Modul za korekciju grafa implementiran je u programskom jeziku C++, a razvoj i testiranje obavljani su u razvojnoj okolini Xcode 4.6.3 na operacijskom sustavu Mac OS X 10.8.4 Mountain Lion. Tijekom razvoja modula, korišten je sustav za verzioniranje GIT 1.8 s online repozitorijem izvornog koda koji je dostupan na adresi:

https://bitbucket.org/Sterbic/graphrefiner



4.1. Algoritmi pretraživanja

Algoritmi pretraživanja prostora stanja ključni su u pronalaženju grešaka u grafu preklapanja. U modulu za korekciju grafa koriste se modifikacije dvaju takvih algoritama koji će biti predstavljeni u ovom potpoglavlju u svojoj općenitoj verziji koja je neovisna o domeni primjene.

Problem pretraživanja definiran je skupom stanja S, početnim stanjem s_0, prijelaza između stanja i ciljnog stanja. Funkcija sljedbenika, $succ : S \rightarrow \varphi(S)$, definira prijelaze između stanja, a funkcija cilja $goal : S \rightarrow \{0, 1\}$ vraća istinu (1) ako je promatrano stanje ciljno stanje, inače vraća neistinu (0). Pretraživanje prostora stanja svodi se na pretraživanje težinskog digrafa koji može biti zadan eksplicitno ili implicitno (eksplicitno kod promatranog modula za korekciju).

Pretraživanjem grafa gradi se stablo pretraživanja tako da se trenutno promatrani čvor proširuje primjenom funkcije $succ$. Tijekom prolaska kroz stablo razlikuju se dva skupa čvorova, skup otvorenih i skup zatvorenih čvorova. Zatvoreni čvorovi su oni čvorovi koji su provjereni za ciljno stanje i prošireni, a otvoreni čvorovi su oni čvorovi koji su generirani iz zatvorenih čvorova, ali nije još primijenjena funkcija sljedbenika nad njima.

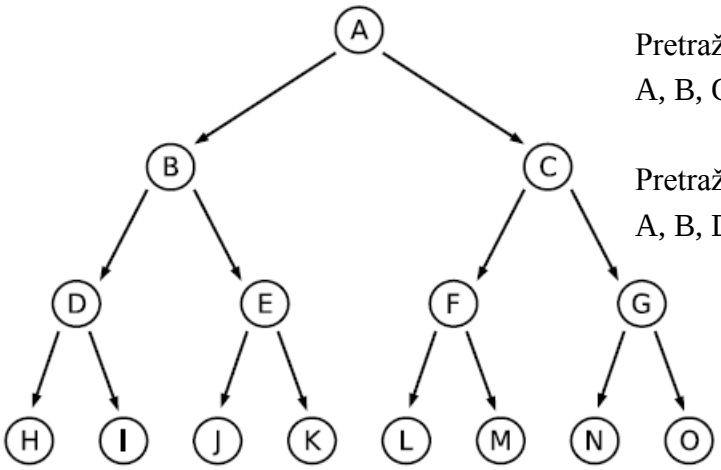
Opći algoritam pretraživanja [8] prikazan je u nastavku. Funkcija izvadiPrvi briše glavu liste i vraća njenu vrijednost. Funkcija ekspandiraj proširuje zadani čvor uporabom funkcije sljedbenika succ. Funkcija umeće zadani čvor u listu.

```
function traži(s0, succ, goal):
 otvoreni ← [početnoStanje(s0)]
 while otvoreni ≠ [] do
      n ← izvadiPrvi(otvoreni)
      if (goal(stanje(n))) then return n
      for m ∈ ekspandiraj(n, succ) do
           ubaci(open, m)
 return greška
```

Algoritmi pretraživanja razlikuju se po načinu obilaska čvorova, tj. redoslijedu kojim se čvorovi proširuju i ubacuju. U izrađenom modulu koriste se prvenstveno dva takva algoritma, pretraživanje u dubinu i pretraživanje u širinu.

Pretraživanje u dubinu (*eng. depth-first search, DFS*) umeće novo otvorene čvorove na početak liste otvorenih čvorova, tj. koristi spomenutu listu kao stog i uvijek proširuje čvor koji je najdublji u stablu. Ovakvo pretraživanje vremenske je složenosti $O(bm)$ i prostorne složenosti $O(bm)$, gdje je b maksimalni faktor grananja čvorova u stablu pretraživanja, a m je maksimalna dubina stabla. Algoritam nije optimalan ni potpun, ali često se koristi zbog relativno male prostorne složenosti, pogotovo kod rekurzivne verzije algoritme koji postiže prostornu složenost $O(m)$. Kako bi se spriječio problem beskonačne petlje ne dopušta se algoritmu da više puta posjećuje isto stanje. U tu svrhu gradi se lista posjećenih ili zatvorenih stanja. Za svaki čvor, nakon što se provjeri da li je stanje konačno stanje, stanje se zapisuje u spomenutu listu. Prilikom ekspandiranja čvorova samo čvorovi sa stanjima koji nisu još posječeni dodaju se u listu otvorenih čvorova.

Pretraživanje u širinu (*engl. breadth-first search, BFS*) ekspandirane čvorove umeće na kraj liste otvorenih čvorova, što rezultira ponašanjem liste otvorenih čvorova kao reda i obradom stabla pretraživanja razinu po razinu. Vremenska i prostorna složenost ovog algoritma jesu $O(b^{d+1})$, gdje je b maksimalni faktor grananja čvorova u grafu, a d je dubina na kojoj se nalazi optimalno rješenje pretraživanja. Algoritam je potpun i optimalan, ali nije primijenjiv na velike probleme zbog eksponencijalne prostorne složenosti. Korištenjem liste posjećenih čvorova složenosti mogu se svesti na $O(\min(b^{d+1}, b|S|)$, gdje je $|S|$ veličina prostora stanja.



Pretraživanje u širinu:
A, B, C, D, E, F, G, H, I, J, K, L, M, N, O

Pretraživanje u dubinu:
A, B, D, H, I, E, J, K, C, F, L, M, G, N, O

Slika 21. Primjer redoslijeda obilaska stabla kod različitih algoritama pretraživanja [8]

4.2. Struktura grafa

Graf preklapanja implementiran je klasom Graph koja u programskom jeziku C++ ima sljedeću strukturu:

```
class Graph {
private:
    int nodeCount;
    int edgeCount;

    bool labeled;
    int maxLabel;

    vector<int> *in;
    vector<int> *out;

    pair<int, int> *edges;
    vector<int> edgeWeights;
    pair<int, int> illegalEdge;

    vector<bool> confirmedNodes;
    vector<bool> confirmedEdges;

    set<int> ban;
    bool foundDirect;
    vector<int> path;

    vector<int> groupLabels;
}
```

Cijeli brojevi *nodeCount* i *edgeCount* predstavljaju ukupni broj čvorova i ukupni broj bridova u grafu. Vrijednost Booleove varijable *labeled* označava da li je graf označen, tj. da li je provedena identifikacija otoka. Cijeli broj *maxLabel* pohranjuje najveću vrijednost oznake koja je korištena prilikom identifikacije otoka počevši od 1.

Varijable *in* i *out* pokazivači su na polje vektora veličine ukupnog broja čvorova u grafu i efektivno se ponašaju kao 2D polje. Svaki element polja je vektor cijelih brojeva koji sadrži indekse zapisa bridova u vektoru bridova koji ulaze ili izlaze iz dotičnog čvora. Npr. izraz `out[7][2]` evaluirao bi se kao indeks trećeg brida koji izlazi iz osmog čvora (indeksiranje kreće od 0). Indeksi bridova referenciraju vrijednosti u vektorima *edges* i *edgeWeights*.

Polje parova cijelih brojeva *edges* veličine je ukupnog broja bridova u grafu i sadrži zapise u obliku para izvorišnog i odredišnog čvora brida. Npr. izraz `edges[3].second` označava izvorišni čvor brida koji je četvrti po redu u vektoru bridova. Vektor *edgesWeights* iste je veličine kao i vektor bridova i referencira se istim indeksima, a svaki zapis označava težinu brida na tom indeksu.

Vektori Booleovih vrijednosti *confirmedEdges* i *confirmedNodes* označavaju je li određeni brid ili čvor potvrđen nakon postupka korekcije grafa. Vektor *confirmedEdges* veličine je ukupnog broja bridova i inicijalno je postavljen na neistinu, dok je vektor *confirmedNodes* veličine ukupnog broja čvorova u grafu i inicijalno je postavljen na istinu.

Vektor cijelih brojeva *groupLabels* koristi se za obilježavanje pripadnosti određenog čvora pojedinom otoku. Ostale varijable, *ban*, *foundDirect* i *path*, koriste se u postupku brisanja lažnih preklapanja.

4.3. Brisanje lažnih preklapanja

Identifikacija i brisanje lažnih bridova koji upućuju na asocijaciju genomskih regija koje u stvarnosti nisu povezane najkompleksniji je dio modula za korekciju grafa. Cilj ovog koraka je dokazati višestruku asocijaciju svih genomskih regija za koje se tvrdi da su povezane, uz pretpostavku normalne razdiobe izmjerenih genomskih udaljenosti.

Za svaki par čvorova N_i i N_j gdje je $i \neq j$ odrediti nezavisne puteve između ta dva čvora. Nezavisni putevi su oni putevi koji nemaju zajedničkih točaka, osim početne i konačne. Određeni parovi čvorova mogu se *a priori* ignorirati ako se uzme u obzir da je broj nezavisnih čvorova ograničen odozgo izrazom::

$$maxPahts = \min(broj\_izlaznih\_bridova(N_i), broj\_ulaznih\_bridova(N_j))$$

ako je broj teoretski mogućih puteva manji od 2, nije potrebna daljnja analiza promatranog para čvorova radi cilja ovog koraka.

Nezavisni putevi između para čvorova grade se iterativnim postupkom koji traje dokle god je u prethodnom koraku pronađen novi put. U ovu svrhu koristi se pretraživanje u dubinu koje uzima u obzir samo izlazne bridove iz svakog čvora i samo čvorove koji nisu već iskorišteni u nekoj prethodnoj iteraciji, tj. koji nisu u skupu zabranjenih čvorova. Kad je put pronađen, svi čvorovi puta osim početnog i konačnog, dodaju se u skup zabranjenih čvorova (*set<int> ban*). Skup zabranjenih čvorova povećava se svakom iteracijom dok se ne iscrpe sve moguće rute.

Ako su barem dva nezavisna puta pronađena, vektor puteva predaje se na daljnju analizu. Za svaki put računa se ukupna genomska udaljenost D_α koju pokriva tako da se sumiraju težine svih bridova kojima prolazi put. Duljina puteva podliježe normalnoj distribuciji

$X \approx N(Y, Y \cdot \sigma^2)$, gdje je X izračunata udaljenost, a Y realna udaljenost. Za svaki put u vektoru puteva pokušava se izgraditi što veći klaster puteva takav da je dužina puteva unutar $\sigma \cdot \sqrt{D_\alpha}$ od dužine puta za kojeg se gradi klaster.

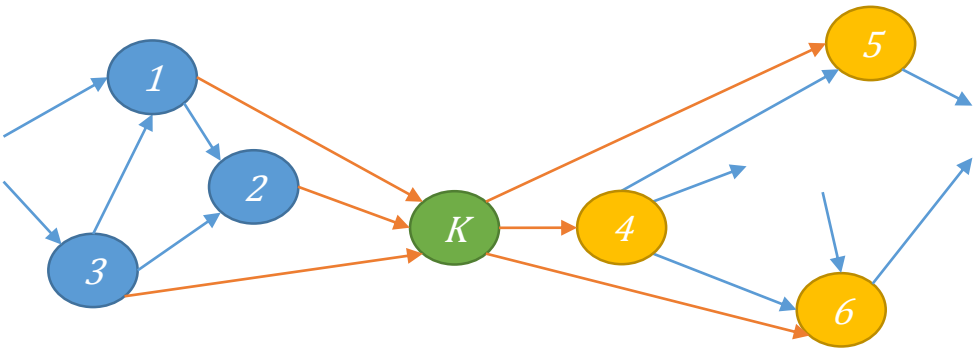
Najveći klaster koji je pronađen u prethodnom koraku smatra se valjanim ako sadrži barem dva nezavisna puta. Svi bridovi puteva takvoga klastera označavaju se kao potvrđeni. Svi bridovi koji nisu potvrđeni ovim postupkom mogu se smatrati lažnima i mogu se obrisati, uz brisanje svih izoliranih čvorova. Izolirani čvor je onaj čvor koji nema ulaznih ni izlaznih bridova ili onaj koji nema niti jedan potvrđeni brid.

4.4. Brisanje kimernih mapa

Kimerne mape nastaju zbog fizičkog preklapanja dvaju mapa tijekom očitavanja i imaju specifičnu topologiju (vidi Slika 20. Primjer kimerne mape (K)). Kimerna mapa uvodi se u graf preklapanja kao jedna mapa iako se u stvarnosti sastoji od dvije odojene mape. Takve greške nije moguće kompletno ispraviti te se zbog toga čvorovi koji su identificirani kao kimerni brišu iz grafa, uz brisanje svih pripadajućih ulaznih i izlaznih bridova.

Procedura eliminacije kimernih mapa iterira po svim čvorovima grafa i za svakog ispituje je li kimeran. Kimernost čvora može se dokazati tako da brisanjem tog čvora graf postaje lokalno nepovezan.

Iz svakog čvora (npr. čvor 3) koji je susjedan čvoru koji se provjerava pokreće se pretraživanje u širinu s ciljem pronalaženja svih ostalih susjeda (čvorovi 1, 2, 4, 5, 6) bez razmatranja puteva koji prolaze kroz mogući kimerni čvor (čvor K). U ovom koraku usmjerenost bridova nije relevantna te se graf tretira kao da nije usmjeren. Ako svi susjedni čvorovi nisu dostižljivi barem iz jednoga čvora, tada se promatrani čvor proglašava kimernim i briše se. Prije puštanja grafa u daljnju obradu, brišu se svi izolirani čvorovi.



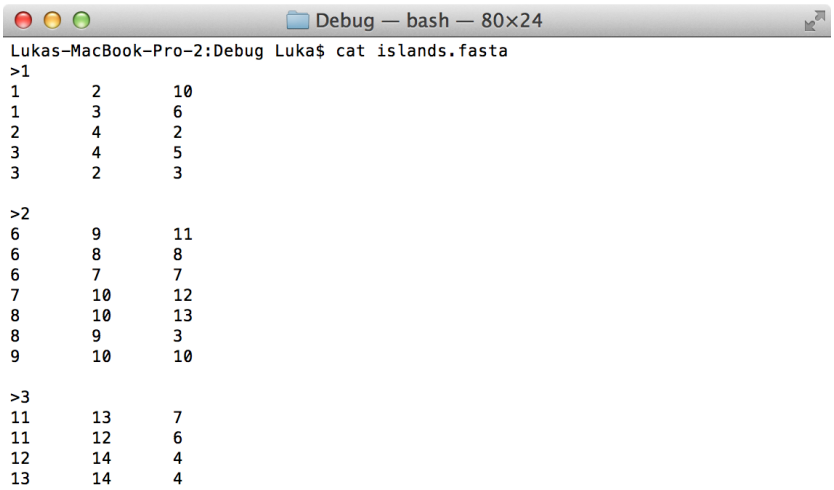
Slika 22. Primjer ispitivanja čvora za kimernost

4.5. Identifikacija otoka

Identifikacija otoka započinje označavanjem svakoga čvora određenom labelom koja se sprema u vektor cijelih brojeva koji je veličine ukupnog broja bridova. Funkcija označavanja iterira po svim čvorovima i za svaki potvrđeni čvor koji nije prethodno označen pokreće proceduru označavanja.

Označavanje pojedinog otoka koristi iscrpno pretraživanje u dubinu kako bi se identificirali svi čvorovi koji pripadaju otoku. Svaki čvor nakon skidanja s liste otvorenih čvorova označava se prethodno definiranom labelom koja se uvećava za jedan nakon obrade pojedinog otoka. Koristi se skup zatvorenih čvorova kako bi se izbjegla mogućnost beskonačne petlje. U ovom koraku usmjerenost bridova nije bitna te se s toga u funkciji ekspandiranja čvorova razmatraju ulazni i izlazni bridovi čvora.

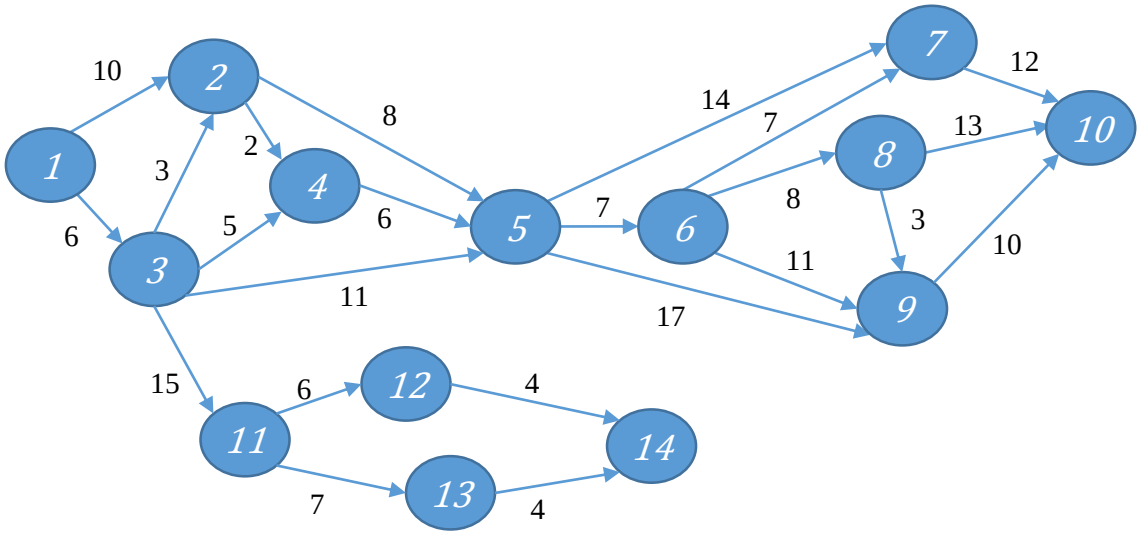
Funkcija koja zapisuje otoke na disk u FASTA formatu iterira po svim labelama (od 1 do *maxLabel*) i svim čvorovima koji su označeni tom labelom. Prvo se ispisuje identifikator otoka u obliku „>L“, gdje je L labela otoka, a zatim se za svaki čvor zapisuje niz redaka koji opisuju izlazne bridove tog čvora. Svaki red sadrži tri cijela broja odvojena tabulatorima: izvorni čvor brida, odredišni čvor brida i težinu brida. Dva susjedna otoka razdvajaju se praznim retkom.



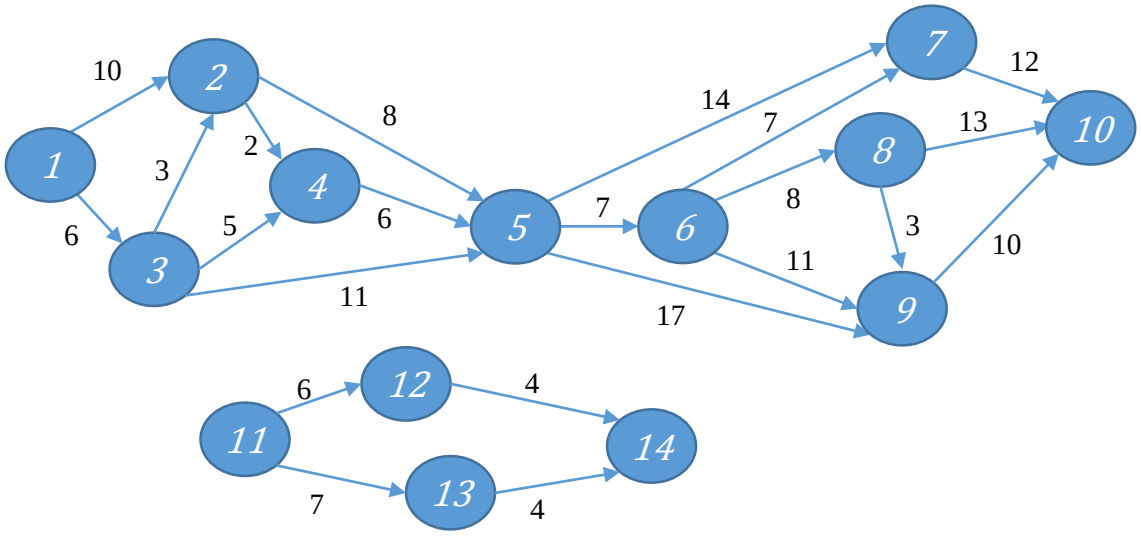
Slika 23. Primjer FASTA datoteke sa zapisom otoka nastalih korekcijom grafa

4.6. Primjer pokretanja programa

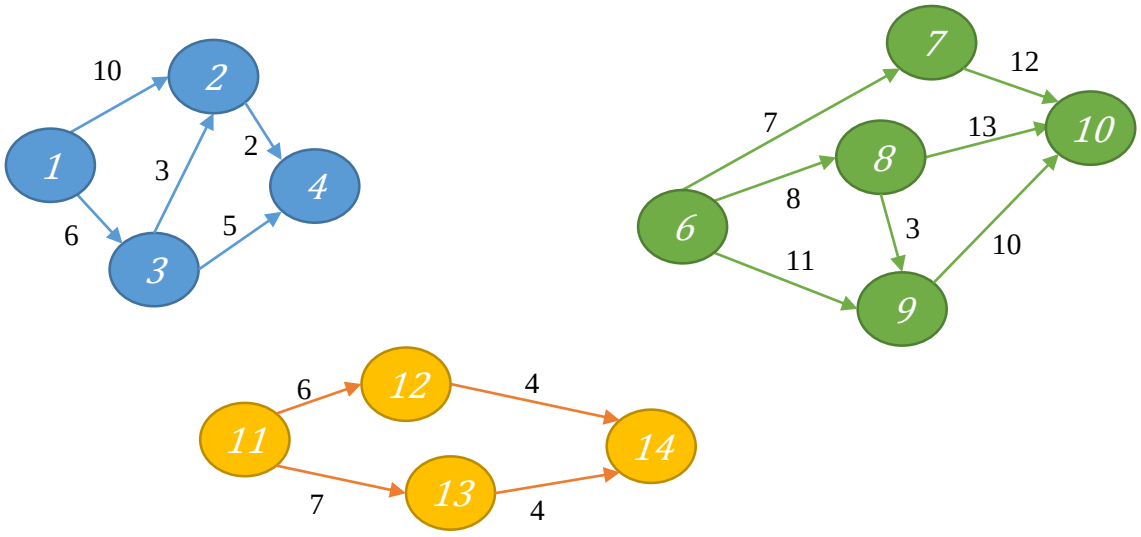
Do izvršne datoteke modula dolazi se pokretanjem naredbe *make* u unix ljusci koja će pokrenuti prevođenje izvornog koda. Program se pokreće naredbom *./graphRefiner* uz navođenje dva argumenta, staze do FASTA datoteke s definicijom grafa i staze do FASTA datoteke u kojoj će biti zapisan izlaz programa. U slikama 24, 25 i 26 prikazan je primjer grafa preklapanja kroz pojedine korake algoritma.



Slika 24. Primjer ulaznog grafa preklapanja



Slika 25. Graf preklapanja nakon eliminacije lažnih bridova



Slika 26. Graf preklapanja nakon eliminacije kimernih čvorova i identifikacije otoka

Radi lakše analize rezultata izrađena je skripta koja uspoređuje ulazni graf i otoke dobivene modulom za korekciju grafa. Skripta je napisana u programskom jeziku Python 3 i prima dva argumenta, putanju do datoteke s definicijom grafa i putanju do datoteke s definicijom otoka. Skripta ispisuje podatke o grafu i otocima, te njihovu razliku u smislu ukupnog broja čvorova i bridova. Za prikazani primjer vidljivo je da je modul za korekciju grafa izbrisao jedan čvor i deset bridova.

```
graph_refiner — bash — 80×24
Lukas-MacBook-Pro-2:graph_refiner Luka$ ./src/islandAnalyzer.py graphs/graph_1.f
asta results/islands_1.fasta
##################################
Input graph file: graphs/graph_1.fasta
Graph stats:
        >|V| = 14
        >|E| = 23
##################################
Input islands file: results/islands_1.fasta
Islands stats:
        >|I| = 3
        >|V| = 13
        >|E| = 16
        >dV = -1
        >dE = -10
##################################
Lukas-MacBook-Pro-2:graph_refiner Luka$
```

Slika 27. Primjer pozivanja skripte za analizu rezultata

Zaključak

Problem asembliranja optičkih mapa pokazao se izrazito zahtjevnim jer optičko mapiranje koristi mjerenja pojedinačnih molekula. Za neke formulacije je dokazano da se radi od NP–teškom problemu, dok se za druge pokazalo mogućnost korištenja algoritma s polinomijalnim složenostima. Takvi algoritmi pokazali su se preciznima za „male“ genome koji su dobiveni kloniranjem, ali nisu prikladni za veće genome i ulazne podatke koji proizlaze iz slučajno isjeckanih molekula DNA, tzv. *shotgun* metoda. Takva situacija dovela je do razvoja algoritma koji će uspješno obrađivati genome viših eukariota poput biljaka ili sisavca.

Prvi algoritam za *de novo shotgun* asembliranje na razini cijeloga genoma predstavili su Valuev et al. 2006. godine u radu „An algorithm for assembly of ordered restriction maps from single DNA molecules“. Njihov algoritam koristi trorazinski okvir overlap – layout – consensus preuzet iz sekvencijskih asemblera. Značajna preklapanja među mapama predstavljaju se kao bridovi težinskoga usmjerenoga grafa, a ključni korak algoritma je primjena izrazito efikasnog mehanizma korekcije pogrešaka u vidu lažnih preklapanja i kimernih mapa. Nakon korekcije graf se raspada na otoke koje predstavljaju neovisne regije genoma, npr. različite kromosome, iz kojih se stvara skica konačne mape te regije. Svaka se skica još jednom rafinira za eliminaciju mogućih netočnosti te se isporučuje kao suglasna optička mapa jedne regije genoma.

Dobre osobine i visoki stupanj preciznosti ovog algoritma potvrđene su usporedbom dobivenih optičkih mapa s poznatom sekvencom analiziranih genoma. Korištena struktura preklapanja omogućuje asembliranje i za polimorfne regije genoma koje se nalaze u diploidnim organizmima i u tumornim stanicama s genomom koji pokazuje jake aberacije.

Glavno područje primjene ovog algoritma bit će strukturalna genomika u kojoj restrikcijske mape mogu otkriti kilobazne alteracije u promatranim genomima kao nova restrikcijska mjesta, nepostojeća restrikcijska mjesta, insercije ili delecije veće od 5 kb i složena restrukturiranja. Optičko mapiranje već je uspješno iskorišteno u analizi gena BRCA1/2, ljudskog kromosoma 22, Beckwith-Wiedman lokusa i ljudskog mitohondrijalnoga genoma.

Literatura

[1] VALOUEV , A; SCHWARTZ, D.C; ZHOU, S; WATERMAN, M.S: "An algorithm for assembly of ordered restriction maps from single DNA molecules", *Pnas vol. 103 no.43*, http://www.pnas.org/content/103/43/15770, listopad 2006.

[2] YOUNG, W.Y; LOCKE, J.C.W; ALTINOK, A; ROSENFELD, N; BACARIAN, T; SWAIN, P.S; MJOLSNESS, E; ELOWITZ, M.B: "Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy", *Nature Protocols 7*, http://www.nature.com/nprot/journal/v7/n1/full/nprot.2011.432.html, 2012.

[3] BROWN, T.A: "Genomes 2", Oxford: Wiley-Liss, dostupno na: http://www.ncbi.nlm.nih.gov/books/NBK21128/, 2002.

[4] RAMME, A. J: "A Review of Optical Mapping as a Method of Whole Genome Analysis", University of Iowa, svibanj 2009.

[5] NOOR, MOHAMED: "Introduction to Genetics and Evolution", Duke University, Coursera 2013, https://www.coursera.org/course/geneticsevolution, snimke i materijali s predavanja

[6] *Human Genome Project, "SNP Fact Sheet"*, s Interneta, http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml

[7] Wiki, "Gel electrophoresis", Wikipedia, s Interneta, http://en.wikipedia.org/wiki/Gel_electrophoresis

[8] DALBELO BAŠIĆ, B; ŠNAJDER, J: "Pretraživanje prostora stanja", FER, http://www.fer.unizg.hr/predmet/umjint, materijali s predavanja, 2013.

[9] YOUNG, W.Y; LOCKE, J.C.W; ALTINOK, A; ROSENFELD, N; BACARIAN, T; SWAIN, P.S; MJOLSNESS, E; ELOWITZ, M.B: "Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy", *Nature Protocols 7*, http://www.nature.com/nprot/journal/v7/n1/full/nprot.2011.432.html, 2012.

[10] Wiki, "Optical mapping", Wikipedia, s Interneta, http://en.wikipedia.org/wiki/Optical_mapping

[11] BROWN, T.A: "Genomes 3", Garland Science, 2007.

Sastavljanje optičkih mapa: modul za korekciju grafa

Sažetak

Optičko mapiranje je metoda mapiranja restrikcijskih lokacija pojedinačne molekule DNA koju su predložili Schwartz et al. 1995. godine. Takve mape mogu otkriti insercije, delecije, inverzije i ponavljanje genetskog materijala te služe za uspostavu korelacije između genotipa i fenotipa u kliničkoj medicini. U ovom je radu predstavljen algoritam za *de novo shotgun* sastavljanje optičkih mapa na razini cjelokupnoga genoma kojeg su predložili Valuev et al. 2006. godine. Problem sastavljanja mape reprezentiran je kao usmjereni težinski graf čiji su čvorovi pojedine optičke mape, a bridovi predstavljaju preklapanja dviju mapa. Takav graf, kad je tek konstruiran, sadrži pogreške u smislu lažnih čvorova (kimerne mape) i lažnih bridova (lažna preklapanja). Bez brisanja lažnih elemenata konstrukcija, suglasna mapa ne bi bila moguća. Korekcija grafa odvija se kroz dvije faze i rezultira raspadom grafa na izolirane komponente, tzv. otoke. Prva faza koristi pretraživanje u dubinu kako bi se ustanovila višestruka asocijacija pojedinih genomskih regija, a druga koristi pretraživanje u širinu za identifikaciju kimernih čvorova.

Ključne riječi: restrikcijsko mapiranje, optičko mapiranje, sastavljanje mapa, teorija grafova, algoritmi pretraživanja

Optical map assembly: graph correction module

Abstract

Optical mapping is a technique for generating ordered restriction maps for single DNA molecules introduced by Schwartz et al. in 1995. This type of maps can identify insertions, deletions, inversions and repeats of genetic material and are used in clinical medicine to establish genotype-phenotype correlation. In this paper, an algorithm for de novo shotgun optical map assembly is described. This genome wide assembly algorithm was first introduced by Valuev et al. in 2006. The assembly problem is described with a weighted directed graph, where nodes represent individual optical maps while edges represent overlaps between maps. Such a graph, after the construction step, contains errors in the form of spurious nodes (chimeric maps) and spurious edges (false overlaps). Without an error correcting step, the construction of a consensus map is usually not possible. The graph correction procedure is carried out in two steps and results in the breakdown of the original graph to isolated components, also called islands. In the first step, a depth-first search is performed to establish multiple connectivity between two genomic regions, while the second step uses a breadth-first search to identify chimeric nodes.

Keywords: restriction mapping, optical mapping, map assembly, graph theory, search algorithms